Fen-Edebiyat Fakültesi	BÜTÜNLEME SINAV KAĞIDI	
Adı:	**Dersin Adı: REGRESYON ANALİZİ**	**Not**
Soyadı:	**Dersin Kodu: IST3011**	
Numarası:	**Bölümü: İSTATİSTİK**	
İmzası:	**Sınav Tarihi: 18/02/2021 Saat 21:00-23:10**	

Açıklamalar

1. A4 biçiminde olan **cevap kağıdınızın her birine ad, soyad, okul numarası yazınız ve imza atınız.**
2. Sınav ile ilgili problemleriniz için sınav süresince **fatih.kizilaslan@marmara.edu.tr** e-posta adresinden iletişime geçebilirsiniz.
3. Türkçe haricinde açıklamalar, karalama biçiminde olan yazılar, nereden geldiği belli olmayan tüm ifadeler cevap olarak kabul edilmeyecektir.
   Açıklaması olmayan cevaplar değerlendirilmeyecektir.
4. Cevaplarınızı anlaşılır ve okunabilecek bir biçimde sisteme yükleyiniz.
5. **Bu sınava katılan her öğrenci bu kuralları ve önceden ilan edilmiş tüm kuralları kabul etmiş olarak değerlendirilecektir.**

SINAV İLE İLGİLİ AÇIKLAMALAR

Cevaplarınızı R Markdown kullanarak oluşturunuz. Yazmanız gereken matematiksel ifadeleri soru numarasını yazarak A4 kağıdına yazabilirsiniz. Oluşturduğunuz R Markdown ve A4 kağıdındaki çözümlerinizi birleştirerek PDF formatında sisteme yükleyiniz.

Sınav sonunda ilgili **R Markdown kodunuzu**n adını **"isim_soyisim"** olarak kaydederek **e-posta** ile **fatih.kizilaslan@marmara.edu.tr** adresine gönderiniz.

Soru A (70 puan)

Her bir soruyu R Markdown'da CEVAP NUMARASI ile yazınız. Sadece sorularda sizden istenilenleri açık ve en kısa bir biçimde açıklayınız. Verinin tamamını KESİNLİKLE cevaplarınızda yazdırmayınız. SAYFA SAYINIZI KONTROL EDİNİZ.

Kaggle'da "https://www.kaggle.com/harlfoxem/housesalesprediction" adresinde yer alan (ayrıca sınavdan bir kaç dakika önce BYS'de bulunan **e-posta adreslerinize ve UES sistemi üzerinden gönderdiğim** "kc_house_data.csv") King County (Washington, USA)'de 2014 Mayıs ve 2015 Mayıs ayları arasında satılan evlerin bazı özellikleri ile fiyatlarından oluşan veriyi kullanarak aşağıdaki soruları cevaplayınız. **Bu analiz için anlamlılık düzeyi** $\alpha = 0.05$ **olarak alınacaktır.**

Bu veri toplam 21613 gözlem ve 21 değişkenden oluşmaktadır. **Analizde SADECE aşağıdaki değişkenler kullanılacaktır.**

price	sqft_living	sqft_above	yr_built	bedrooms	bathrooms	waterfront	view	condition	grade
evin fiyatı (USD)	evin yaşam alanı (square feet)	evin giriş üstündeki alanı(square feet)	evin yapım yılı	evin yatak odası sayısı	evin banyo sayısı	evin deniz görme durmu **2 kategori**	evin görünümü **5 kategori**	evin durumu **5 kategori**	ev ile ilgili bir indeks

Not: Bir evin **Grade** indeksi 1-3: ise inşaat ve tasarım **yetersiz**; 7 ise inşaat ve tasarım **ortalama**; 11-13 inşaat ve tasarım **iyi kalite** olduğu anlamına gelmektedir.

1. **(8 puan)** Okul numaranızın **6. basamağındaki rakam** a ve **son iki basamağındaki sayı** b olarak alarak **kc_house_data.csv** verisinin ilk $1250 + [100 * (a + b)]$ gözlemini kullanarak "**my_data**" adında data.frame oluşturunuz.

   Örneğin, okul numaranız 121507085 ise $a = 7$ ve $b = 85$ olmak üzere **kc_house_data.csv** verisinin ilk $1250 + [100 * (85 + 7)] = 10450$ gözlemi ile **my_data** oluşturulur.

   # Aşağıdaki tüm analizler **my_data** verisi için yapılacaktır.

   Yukarıda verilen **kategorik değişkenleri gösterge (dummy) değişken** olarak tanımlayınız.

2. **(12 puan)** my_data verisindeki **nümerik değişkenler** için korelasyon matrisini hesaplayınız ve görselleştiriniz (istediğiniz paketi ve fonksiyonu kullanabilirsiniz).

   price bağımlı değişkeni ile diğer değişkenler arasındaki korelasyona bakarak bu değişkenlerden hangilerini doğrusal regresyon modelinde kullanmak uygun olur? Kısaca açıklayınız.

3. **(15 puan) Model_1: price** bağımlı değişken ve **sqft_living, sqft_above, yr_built, bedrooms, bathrooms, waterfront, view, condition, grade** bağımsız değişkenler olmak üzere çoklu doğrusal regresyon modelini oluşturunuz.

   Model_1 anlamlı mıdır?

   Bağımsız değişkenlerin anlamlılıkları için ne söylenebilir?

   R^2 ve R^2_{adj} değerlerini yorumlayınız.

   (Anlaşılır bir biçimde kısaca açıklayınız.)

4. **(5 puan)** Model_1'den **sqft_above** ve **condition** değişkenlerini çıkararak **Model_2** oluşturunuz. Model_2 anlamlı mıdır? Açıklayınız

5. **(5 puan) sqft_above ve condition** değişkenlerinin anlamlılığını kısmi F testi ile test ediniz (sadece R programını kullanarak). Sonuçlarını açıklayınız.

6. **(10 puan)** R^2, R^2_{adj} değerleri ve kısmi F testinin sonucuna göre Model_1 ve Model_2'den hangisini tercih edersiniz? Kısaca açıklayınız.

7. **(15 puan)** Yaşadığınız ev için **sqft_living, yr_built, bedrooms, bathrooms, waterfront** değerlerini oluşturunuz.

   Not: sqft_living modelde "square feet" birimindedir. $1m^2 = 10.76$ square feet dönüşümünü kullanınız.

   a) Kendi değerleriniz ile birlikte **view=2** ve **grade=7** değerlerini kullanarak **price** yanıt değişkeninin tahmin değerini bulunuz.

   b) Kendi değerleriniz ile birlikte **view=3** ve **grade=7** değerlerini kullanarak **price** yanıt değişkeninin tahmin değerini bulunuz.

   Bulduğunuz bu iki tahmin değeri arasındaki fark ne ile ilşikilidir? Açıklayınız.

Soru B (30 puan)

(60+b) tane gözlem ve 12 tane bağımsız değişken kullanılarak bir çoklu doğrusal regresyon modeli oluşturulmuştur. Bu model için varyansın yansız tahmini $\widehat{\sigma}^2 = 10$ ve $R^2 = 0.92$ olarak hesaplanmıştır. Bu verilenlere göre aşağıdaki soruları **anlamlılık düzeyi** $\alpha = 0.01$ olmak üzere cevaplayınız. **(Not: t ve F tablo değerlerini R programı ile hesaplayınız.)**

1. **(15 puan)** Bu tam (full) model için ANOVA tablosunu oluşturunuz. Regresyon modelinin anlamlılığı için gerekli hipotezleri yazınız ve test ediniz.

2. **(15 puan)** Tam modelden ilk a tane bağımsız değişken $(x_1, x_2, ..., x_a)$ çıkartılarak indirgenmiş bir model oluşturulmuş ve bu yeni model için $SSE = 1375$ olarak hesaplanmıştır.

   Bu durumda bu çıkarılan değişkenlerin anlamlılığı için gerekli hipotezleri yazarak anlamlılıklarını test ediniz. Bulduğunuz sonucu **en fazla 2 cümle** ile açıklayınız.

BAŞARILAR **Doç. Dr. Fatih KIZILASLAN**

IST3011 2020-2021 Guz Bütünleme Sınavı Cevap Anahtarı

Fatih Kızılaslan

10 12 2020

İçindekiler

Soru A (70 puan) **1**

1 (8 puan) . . . 1

2 (12 puan) . . . 3

3 (15 puan) . . . 4

4 (5 puan) . . . 5

5 (5 puan) . . . 6

6 (10 puan) . . . 7

7 (15 puan) . . . 7

Soru B (30 puan) **8**

1 (15 puan) . . . 9

2 (15 puan) . . . 9

Soru A (70 puan)

1 (8 puan)

Verinin excelden alınması ve **my_data**'nın oluşturulması.

```
house_data<- read.csv("kc_house_data.csv", header=TRUE)
attach(house_data)
head(house_data)
```

```
##           id            date   price bedrooms bathrooms sqft_living sqft_lot
## 1 7129300520 20141013T000000  221900        3      1.00        1180     5650
## 2 6414100192 20141209T000000  538000        3      2.25        2570     7242
## 3 5631500400 20150225T000000  180000        2      1.00         770    10000
## 4 2487200875 20141209T000000  604000        4      3.00        1960     5000
## 5 1954400510 20150218T000000  510000        3      2.00        1680     8080
## 6 7237550310 20140512T000000 1225000        4      4.50        5420   101930
##   floors waterfront view condition grade sqft_above sqft_basement yr_built
## 1      1          0    0         3     7       1180             0     1955
```

```
## 2      2          0    0         3     7       2170           400     1951
## 3      1          0    0         3     6        770             0     1933
## 4      1          0    0         5     7       1050           910     1965
## 5      1          0    0         3     8       1680             0     1987
## 6      1          0    0         3    11       3890          1530     2001
##   yr_renovated zipcode     lat     long sqft_living15 sqft_lot15
## 1            0   98178 47.5112 -122.257          1340       5650
## 2         1991   98125 47.7210 -122.319          1690       7639
## 3            0   98028 47.7379 -122.233          2720       8062
## 4            0   98136 47.5208 -122.393          1360       5000
## 5            0   98074 47.6168 -122.045          1800       7503
## 6            0   98053 47.6561 -122.005          4760     101930
```

```
a=7
b=85
1250+(100*(a+b))
```

```
## [1] 10450
```

```
my_data<-data.frame(house_data[1:(1250+(100*(a+b))),])
head(my_data)
```

```
##           id            date   price bedrooms bathrooms sqft_living sqft_lot
## 1 7129300520 20141013T000000  221900        3      1.00        1180     5650
## 2 6414100192 20141209T000000  538000        3      2.25        2570     7242
## 3 5631500400 20150225T000000  180000        2      1.00         770    10000
## 4 2487200875 20141209T000000  604000        4      3.00        1960     5000
## 5 1954400510 20150218T000000  510000        3      2.00        1680     8080
## 6 7237550310 20140512T000000 1225000        4      4.50        5420   101930
##   floors waterfront view condition grade sqft_above sqft_basement yr_built
## 1      1          0    0         3     7       1180             0     1955
## 2      2          0    0         3     7       2170           400     1951
## 3      1          0    0         3     6        770             0     1933
## 4      1          0    0         5     7       1050           910     1965
## 5      1          0    0         3     8       1680             0     1987
## 6      1          0    0         3    11       3890          1530     2001
##   yr_renovated zipcode     lat     long sqft_living15 sqft_lot15
## 1            0   98178 47.5112 -122.257          1340       5650
## 2         1991   98125 47.7210 -122.319          1690       7639
## 3            0   98028 47.7379 -122.233          2720       8062
## 4            0   98136 47.5208 -122.393          1360       5000
## 5            0   98074 47.6168 -122.045          1800       7503
## 6            0   98053 47.6561 -122.005          4760     101930
```

Nicel değişkenler:

bağımlı değişken: price bağımsız değişkenler: sqft living, sqft above, yr built, bedrooms, bathrooms, grade

Gösterge değişkenlerin oluşturulması.

```
my_data$waterfront<-as.factor(my_data$waterfront)
my_data$view<-as.factor(my_data$view)
my_data$condition<-as.factor(my_data$condition)

str(my_data)
```

```
## 'data.frame':    10450 obs. of  21 variables:
##  $ id           : num  7.13e+09 6.41e+09 5.63e+09 2.49e+09 1.95e+09 ...
##  $ date         : chr  "20141013T000000" "20141209T000000" "20150225T000000" "20141209T000000" ...
##  $ price        : num  221900 538000 180000 604000 510000 ...
##  $ bedrooms     : int  3 3 2 4 3 4 3 3 3 3 ...
##  $ bathrooms    : num  1 2.25 1 3 2 4.5 2.25 1.5 1 2.5 ...
##  $ sqft_living  : int  1180 2570 770 1960 1680 5420 1715 1060 1780 1890 ...
##  $ sqft_lot     : int  5650 7242 10000 5000 8080 101930 6819 9711 7470 6560 ...
##  $ floors       : num  1 2 1 1 1 1 2 1 1 2 ...
##  $ waterfront   : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ view         : Factor w/ 5 levels "0","1","2","3",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ condition    : Factor w/ 5 levels "1","2","3","4",..: 3 3 3 5 3 3 3 3 3 3 ...
##  $ grade        : int  7 7 6 7 8 11 7 7 7 7 ...
##  $ sqft_above   : int  1180 2170 770 1050 1680 3890 1715 1060 1050 1890 ...
##  $ sqft_basement: int  0 400 0 910 0 1530 0 0 730 0 ...
##  $ yr_built     : int  1955 1951 1933 1965 1987 2001 1995 1963 1960 2003 ...
##  $ yr_renovated : int  0 1991 0 0 0 0 0 0 0 0 ...
##  $ zipcode      : int  98178 98125 98028 98136 98074 98053 98003 98198 98146 98038 ...
##  $ lat          : num  47.5 47.7 47.7 47.5 47.6 ...
##  $ long         : num  -122 -122 -122 -122 -122 ...
##  $ sqft_living15: int  1340 1690 2720 1360 1800 4760 2238 1650 1780 2390 ...
##  $ sqft_lot15   : int  5650 7639 8062 5000 7503 101930 6819 9711 8113 7570 ...
```

2 (12 puan)

my_data'da bulunan nicel değişkenler için korealasyon matrisinin oluşturulması.

```
library(corrplot) #package corrplot
```

```
## corrplot 0.84 loaded
```

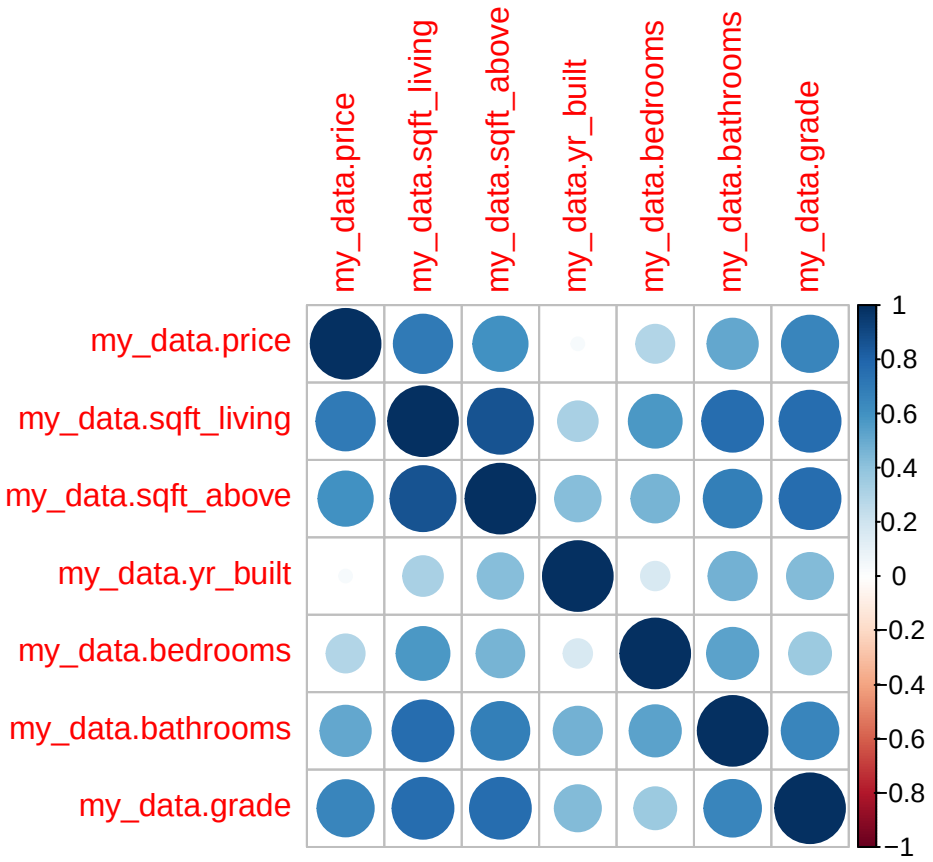
```
my_data_1<-data.frame(my_data$price,my_data$sqft_living,my_data$sqft_above,
    my_data$yr_built,my_data$bedrooms,my_data$bathrooms,my_data$grade)

round(cor(my_data_1),5)
```

```
##                     my_data.price my_data.sqft_living my_data.sqft_above
## my_data.price             1.00000             0.70021            0.60570
## my_data.sqft_living       0.70021             1.00000            0.86901
## my_data.sqft_above        0.60570             0.86901            1.00000
## my_data.yr_built          0.03348             0.32399            0.42532
## my_data.bedrooms          0.29765             0.57468            0.46701
## my_data.bathrooms         0.51974             0.76244            0.68553
## my_data.grade             0.65040             0.76308            0.76030
##                     my_data.yr_built my_data.bedrooms my_data.bathrooms
## my_data.price                0.03348          0.29765           0.51974
## my_data.sqft_living          0.32399          0.57468           0.76244
## my_data.sqft_above           0.42532          0.46701           0.68553
## my_data.yr_built             1.00000          0.16694           0.47873
## my_data.bedrooms             0.16694          1.00000           0.53928
## my_data.bathrooms            0.47873          0.53928           1.00000
## my_data.grade                0.43700          0.36185           0.65975
```

```
##                       my_data.grade
## my_data.price               0.65040
## my_data.sqft_living         0.76308
## my_data.sqft_above          0.76030
## my_data.yr_built            0.43700
## my_data.bedrooms            0.36185
## my_data.bathrooms           0.65975
## my_data.grade               1.00000
```

```
corrplot(cor(my_data_1), method = "circle") #plot matrix
```



Yukarıda bulunan korelasyon matrisinde **price** değişkeni ile **yr_built** arasında doğrusal bir ilişki olmadığını söyleyebiliriz. Diğer değişkenler ile **price** arasında bazılarında zayıf da olsa doğrusal ilişkinin olduğu görülmektedir. Bu nedenle **yr_built** değişken için dikkatli davranarak tüm değişkenleri kullanarak doğrusal regresyon modeli oluşturmak bu aşamada doğrudur. **yr_built** değişkeni için ise modelden elde edilecek sonuçlara bakılmaldır.

3 (15 puan)

price ~ sqft_living + sqft_above + bedrooms + bathrooms + yr_built + grade + view + waterfront + condition için çoklu doğrusal regresyon modeli.

```
Model_1<-lm(price~sqft_living+sqft_above+bedrooms+bathrooms+yr_built+grade+view+
    waterfront+condition,data=my_data)
summary(Model_1)
```

```
##
## Call:
## lm(formula = price ~ sqft_living + sqft_above + bedrooms + bathrooms +
##     yr_built + grade + view + waterfront + condition, data = my_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1383567  -110489    -7421    91119  4151263
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.068e+06  1.979e+05  35.707  < 2e-16 ***
## sqft_living  1.752e+02  6.249e+00  28.043  < 2e-16 ***
## sqft_above   2.011e+01  5.973e+00   3.366 0.000765 ***
## bedrooms    -4.520e+04  3.026e+03 -14.935  < 2e-16 ***
## bathrooms    4.784e+04  4.893e+03   9.779  < 2e-16 ***
## yr_built    -3.954e+03  9.897e+01 -39.955  < 2e-16 ***
## grade        1.172e+05  3.190e+03  36.751  < 2e-16 ***
## view1        8.959e+04  1.693e+04   5.292 1.23e-07 ***
## view2        4.213e+04  1.048e+04   4.019 5.88e-05 ***
## view3        1.120e+05  1.472e+04   7.609 3.00e-14 ***
## view4        2.809e+05  2.321e+04  12.102  < 2e-16 ***
## waterfront1  5.930e+05  3.058e+04  19.392  < 2e-16 ***
## condition2  -1.791e+04  6.800e+04  -0.263 0.792318
## condition3  -9.291e+03  6.394e+04  -0.145 0.884470
## condition4   2.065e+03  6.395e+04   0.032 0.974242
## condition5   3.052e+04  6.424e+04   0.475 0.634748
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 220600 on 10434 degrees of freedom
## Multiple R-squared:  0.6537, Adjusted R-squared:  0.6532
## F-statistic:  1313 on 15 and 10434 DF,  p-value: < 2.2e-16
```

a). Model_1 için p-değeri<0.05 olduğundan oluşturulan regresyon modeli anlamlıdır.

b). Bağımsız değişkenlerin anlamlılıkları için her biri için kısmi t-testlerinin sonuçlarına baklır. Yukarıdaki sonuçlara göre **condition** değişkeni haricinde bulunan tüm bağımsız değişkenler için kısmi t-testlerini sonucu $\alpha = 0.05$'den küçük olduğu için bu değişkenlerin her biri diğer değişkenler modeldeyken anlamlıdır.

condition değişkeni 5 seviyesi olan bir kategorik değişkendir. Bu nedenle bunun için modelde 4 farklı bağımsız değişken vardır (condition2,. . . ,condition5 gibi). Bu değişkenlerin her biri için p-değerleri 0.05'den büyüktür. Bu nedenle **condition** kategorik bağımsız değişkeni diğer değişkenler modeldeyken model anlamlı bir katkısı bulunmamaktadır.

c). $R^2 = 0.6537$ ve $R^2_{Adj} = 0.6532$ bulunmuştur. Dolayısıyla oluşturulan modelde bağımsız değişkenler bağımlı değişken **price**'daki değişimin yaklaşık 'ini açıklamaktadır.

4 (5 puan)

sqft_above ve **condition** değişkenleri çıkarılarak Model_2 oluşturulur.

```
Model_2<-lm(price~sqft_living+bedrooms+bathrooms+yr_built+grade+view+waterfront,data=my_data)
summary(Model_2)
```

```
##
## Call:
## lm(formula = price ~ sqft_living + bedrooms + bathrooms + yr_built +
##     grade + view + waterfront, data = my_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1387404  -111138    -7093    90999  4141660
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.193e+06  1.771e+05  40.617  < 2e-16 ***
## sqft_living  1.892e+02  4.768e+00  39.688  < 2e-16 ***
## bedrooms    -4.526e+04  3.023e+03 -14.971  < 2e-16 ***
## bathrooms    4.844e+04  4.867e+03   9.951  < 2e-16 ***
## yr_built    -4.023e+03  9.319e+01 -43.172  < 2e-16 ***
## grade        1.190e+05  3.090e+03  38.524  < 2e-16 ***
## view1        8.514e+04  1.690e+04   5.038 4.77e-07 ***
## view2        3.918e+04  1.043e+04   3.755 0.000174 ***
## view3        1.063e+05  1.463e+04   7.269 3.89e-13 ***
## view4        2.754e+05  2.314e+04  11.901  < 2e-16 ***
## waterfront1  5.970e+05  3.059e+04  19.512  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 220900 on 10439 degrees of freedom
## Multiple R-squared:  0.6526, Adjusted R-squared:  0.6523
## F-statistic:  1961 on 10 and 10439 DF,  p-value: < 2.2e-16
```

Model_2 için p-değeri<0.05 olduğundan oluşturulan regresyon modeli anlamlıdır.

5 (5 puan)

sqft_above ve **condition** değişkenleri için kısmi F testi aşağıdaki gibi uygulanır.

```
A1<-anova(Model_1)
A2<-anova(Model_2)

anova(Model_1,Model_2)
```

```
## Analysis of Variance Table
##
## Model 1: price ~ sqft_living + sqft_above + bedrooms + bathrooms + yr_built +
##     grade + view + waterfront + condition
## Model 2: price ~ sqft_living + bedrooms + bathrooms + yr_built + grade +
##     view + waterfront
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1  10434 5.0759e+14
## 2  10439 5.0920e+14 -5 -1.605e+12 6.5984 3.859e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
Ftablo<-print(qf(1-0.05,5,10434))
```

```
## [1] 2.214956
```

anova(Model_1,Model_2)'in sonuçlarına göre F istatistiğinin değeri 6.5984 $F_{tablo} = F_{0.05,5,10434} = 2.215$'dan büyük olduğundan çıkarılan değişkenlerin modelde diğer değişkenler varken anlamlı bir katkısı vardır.

Ayrıca, F testinin değerini aşağıdaki gibi de hesaplayabiliriz.

```r
SSR_full<-sum(A1$`Sum Sq`[1:9])
SSE_full<-A1$`Sum Sq`[10]
SSR_reduced<-sum(A2$`Sum Sq`[1:7])

F0_condition<- print( ((SSR_full-SSR_reduced)/5) /(SSE_full/A1$`Df`[10] ) )
```

```
## [1] 6.598445
```

```r
Ftablo<qf(1-0.05,5,A1$`Df`[10])
```

```
## [1] FALSE
```

6 (10 puan)

Model_1 için $R^2 = 0.6537$ ve $R^2_{Adj} = 0.6532$

ve

Model_2 için $R^2 = 0.6526$ ve $R^2_{Adj} = 0.6523$

bulunmuştur. Her iki model için bu değerler birbirlerine oldukça yakındır.

5. sorudaki kısmi F testine göre çıkarılan **sqft_above** ve **condition** değişkenlerinin modele anlamlı bir katkısı vardır. Ancak, Model_1'de kımsi t testine göre **condition** anlamsızdır.

Bu durumda kısmi F testi ve R^2_{Adj} sonucundan dolayı Model_1 tercih edilebilir.

Ayrıca,

1. Model_1'de bulunan **sqft_above** ve **condition** değişkenlerinin maliyetleri gözönünde bulundurularak kullanılmaması da tercih edilebilir.

2. Model_1'de **condition** olmadığı bir model oluşturup sonuçlarını bu modellerle karşılaştırabiliriz.

7 (15 puan)

Yaşadığımız ev için oluşturduğumuz değerler, verilen **view** ve **grade** değerleri için tahmin değerleri Model_2'ye göre hesaplanır.

Örneğin, evimiz $100m^2$, 2000 yılında yapılmış, 2 yatak odası, 1 banyosu, deniz görünümü yok yani waterfront=0 ise sonuçlar aşağıdaki gibidir.

```
pre1<-predict(Model_2,newdata=data.frame(sqft_living=c(100*10.76), yr_built=c(2000),
 bedrooms=c(2), bathrooms=c(1), waterfront="0", view="2",grade=c(7)))

pre2<-predict(Model_2,newdata=data.frame(sqft_living=c(100*10.76), yr_built=c(2000),
   bedrooms=c(2), bathrooms=c(1), waterfront="0", view="3",grade=c(7)))

pre1
```

```
##        1
## 180391.4
```

```
pre2
```

```
##        1
## 247534.5
```

```
pre2-pre1
```

```
##        1
## 67143.11
```

Bulunan iki değer arasındaki fark Model_2'de yer alan **view** gösterge değişkeninin **view2** ile **view3** için regresyon katsayılarının tahminleri arasındaki farkdır.

Aşğıda yukarıdaki farkın nasıl meydana geldiği görülebilir.

```
Model_2$coefficients
```

```
##  (Intercept)  sqft_living     bedrooms    bathrooms     yr_built        grade
## 7192528.2789     189.2213  -45258.7324   48436.3870   -4023.0341  119033.1028
##        view1        view2        view3        view4  waterfront1
##   85138.5768   39178.6679  106321.7804  275427.5753  596955.5678
```

```
Fark<- print(Model_2$coefficients[9]-Model_2$coefficients[8])
```

```
##    view3
## 67143.11
```

Soru B (30 puan)

Aşağıda B'de sorulan sorular ve ANOVA tabloları için gerekli tüm işlemler ve sonuçlar bulunmaktadır. Elde edilenlere göre ANOVA tabloları oluşturulacaktır.

Verilenler:

```
n<- (60+b)
k<-12
p<- (k+1)
sigma_kare<-10
R2_full<-0.92
alfa<-0.01
```

1 (15 puan)

```
SSEfull<- print((n-p)*sigma_kare)
```

```
## [1] 1320
```

SSEfull=(1-0.92)SST olduğundan

```
SST<-print(SSEfull/(1-0.92))
```

```
## [1] 16500
```

bulunur.

Böylece, regresyonun anlamlılığı için F testinin değeri

```
SSRfull<-print((SST-SSEfull))
```

```
## [1] 15180
```

```
Fh<-print(SSRfull/k)/(SSEfull/n-p)
```

```
## [1] 1265
```

```
Ftablo<-print(qf(1-0.01,k,n-p))
```

```
## [1] 2.32219
```

$Fh > Ftablo$ olduğundan

$H_0 : \beta_1 = \beta_2 = \beta_3 = ... = \beta_{12} = 0$

hipotezi reddedilir. Verilen regresyon modeli anlamlıdır.

2 (15 puan)

İlk **a** değişken modelden çıkarılıyor. Bu durumda $SSEreduced = 1375$ hesaplanmıştır. $SSTfull$ değişmediğinden bu durumda

```
SSEreduced<-1375
SSRreduced<-print(SST-SSEreduced)
```

```
## [1] 15125
```

bulunur.

Çıkarılan değişkenler için kısmi F testinin değeri:

```
Fh_reduced<-print( ((SSRfull-SSRreduced)/a) / (SSEfull/(n-p))   )
```

```
## [1] 0.7857143
```

```
Ftablo_reduced<-print( qf(1-0.01,a,n-p))
```

```
## [1] 2.777549
```

Fh_reduced<Ftablo_reduced olduğundan

$H_0 : \beta_1 = \beta_2 = \beta_3 = ... = \beta_a = 0$

hipotezi kabul edilir. Böylece çıkarılan değişkenler diğer değişkenler modeldeyken anlamlı bir katkısı bulunmamaktadır.