HAKAN SONMEZ
Section-B

15.071 Analytics Edge - Homework Assignment#1

Problem-2: Forecasting Jeep Wrangler and Hyundai Elantra Sales

a.) Initially I started my model including all 5 independent variables: "Year", "Unemployment", "WranglerQueries", "CPI.Energy", and "CPI.All". The R^2 value of that model was 0.8751 which was sufficiently high. However, in this model I observed that the "Unemployment" variable has a relatively high p-value (no star) so it does not significantly impact on the output variable: WranglerSales.

I also used the correlation matrix to find correlations between independent variables and observed that Unemployment is a highly correlated to other independent variables in my initial model.

Thus I have removed Unemployment rate from my initial model and created a new model with "Year", "WranglerQueries", "CPI.All", and "CPI". Energy as my independent variables. In this new model all included independent variables have p-values smaller than 0.05 which indicates they are significantly impactful on the output variable, Wrangler Sales. The R output of the model summary is copied on the following page.

Here is the linear regression equation with coefficients for each variable including the intercept:

```
WranglerSales = 4003711.49 -2038.35*(Year) + 267.37*(WranglerQueries)
+448.85*(CPI.All) -37.57*(CPI.Energy)
```

All independent variables have at least one "star" (*) in the model summary which shows high quality for the model. The new R^2 value of my model is 0.8749 which is still relatively high and very close to the original R^2 value (0.8751) found with my initial model of including all 5 variables.

The coefficients in the linear regression equation made sense in terms since we observe that Wrangler sales increase with the Wrangler search queries. More people searching for the car means, more people are interested in buying the car. The wrangler sales are negatively correlated with CPI.Energy index which also makes sense that increasing energy prices (including oil prices) would decrease the sales of Wrangler which is not necessarily known as a fuel-efficient SUV. We also observe that wrangler sales are positively correlated with CPI.All index which makes sense since usually high prices (inflation) happens when people are buying more goods.

Interestingly our model shows that the coefficient of the "Year" variable is negative, which indicates that Wrangler sales decline by year value increasing. This shows a general trend that Wrangler becomes a less popular product to buy as years pass, which can make sense as more people are preferring fuel-efficient cars each year or there are better SUVs introduced to market as years pass.

The summaries of my initial model (model1) and final model (model3) and the correlation matrix is attached on the following page for your reference.

- **Initial Model with all 5 independent variables:**

```
> model1=lm(WranglerSales ~Year +Unemployment +WranglerQueries +CPI.All + CPI.Energy, data=train)
> summary(model1)

Call:
lm(formula = WranglerSales ~ Year + Unemployment + WranglerQueries +
    CPI.All + CPI.Energy, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-2890.6  -847.6   158.7   791.2  3527.6

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     3888421.05 1195757.22   3.252  0.00181 **
Year              -1988.05     602.55  -3.299  0.00156 **
Unemployment        314.92     895.22   0.352  0.72613
WranglerQueries     266.62      24.42  10.919  < 2e-16 ***
CPI.All             507.16     218.17   2.325  0.02318 *
CPI.Energy          -44.80      25.92  -1.728  0.08864 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1362 on 66 degrees of freedom
Multiple R-squared:  0.8751, Adjusted R-squared:  0.8656
F-statistic: 92.48 on 5 and 66 DF,  p-value: < 2.2e-16
```

- **Final Model with variables Year, WranglerQueries, CPI.All ,CPI.Energy :**

```
> model3 = lm(WranglerSales ~ Year  + WranglerQueries + CPI.All + CPI.Energy, data=train)
> summary(model3)
Call:
lm(formula = WranglerSales ~ Year + WranglerQueries + CPI.All +
    CPI.Energy, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-2920.5  -826.7   121.9   814.1  3565.5

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     4003711.49 1142422.03   3.505 0.000820 ***
Year              -2038.35     581.50  -3.505 0.000818 ***
WranglerQueries     267.37      24.16  11.065  < 2e-16 ***
CPI.All             448.85     140.92   3.185 0.002197 **
CPI.Energy          -37.57      15.71  -2.392 0.019578 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1353 on 67 degrees of freedom
Multiple R-squared:  0.8749, Adjusted R-squared:  0.8674
F-statistic: 117.1 on 4 and 67 DF,  p-value: < 2.2e-16
```

- **Correlation Matrix:**

```
> cor(onedata[,c("WranglerSales", "Year", "Unemployment", "WranglerQueries","CPI.All", "CPI.Energy")])
```

```
                WranglerSales       Year Unemployment WranglerQueries    CPI.All CPI.Energy
WranglerSales       1.0000000  0.7625062   -0.7709101       0.8978846  0.7661560 -0.2856677
Year                0.7625062  1.0000000   -0.9850380       0.9035815  0.9527437 -0.4672274
Unemployment       -0.7709101 -0.9850380    1.0000000      -0.8885485 -0.9508470  0.4742905
WranglerQueries     0.8978846  0.9035815   -0.8885485       1.0000000  0.8547637 -0.4398302
CPI.All             0.7661560  0.9527437   -0.9508470       0.8547637  1.0000000 -0.2136247
CPI.Energy         -0.2856677 -0.4672274    0.4742905      -0.4398302 -0.2136247  1.0000000
```

b.) Google Trends:

I observe that there is a seasonality factor in the search trend for Jeep for Wrangler. I observe that the trend of search queries for Wrangler goes up during spring and summer months and it makes a peak during the mid-summer (around mid-july) and then decreases to make a dip around mid-November. Then the trend to starts to rise again starting spring.

This makes intuitive business sense as car sales tend to increase during the summer months. After sales peak in mid-summer, sales tend to drop during autumn. The search trends are in parallel to the sales trends which makes sense intuitively.

c.) Below is the linear regression equation after adding MonthFactor as an independent variable.

```
WranglerSales = (4.192e+06 -(2.151e+03) *(Year) + (1.636e+02)*(WranglerQueries)+
(6.719e+02)  *(CPI.All) -(6.011e+01)*(CPI.Energy) + -2.280e+03*(MonthFactorFebruary)
-2.828e+03*(MonthFactorJanuary) +1.576e+03*(MonthFactorMay)
- 2.804e+03*(MonthFactorNovember) -1.611e+03*(MonthFactorOctober)
- 1.756e+03*(MonthFactorSeptember)
where Monthfactor<month_name> are have binary values (0,1) and only one
MonthFactor<month_name> can have the value of 1 at a given time.
```

The model has a multiple R^2 value of 0.9527 which is significantly higher than the value we found in Part-a. where we did not include the seasonality factor. Thus we can say that adding MonthFactor has improved the quality of our model. The significant variables are the following ones which has * assigned:

“WranglerQueries”, “CPI.All” ,”CPI.Energy”, “Year”, “Monthfactors: January, February, May, September, October, November”

By looking at the coefficients of the months, we can tell that the month of May has positive impact on the Wrangler sales (sales increase) and other significant months, January, February, September, October and November have a negative impact on the Wrangler sales (sales decline).

The resulting linear regression equation is the following:

```
> model12 = lm(WranglerSales ~ Year + MonthFactor + WranglerQueries + CPI.All + CPI.Energy,
data=train)
> summary(model12)

Call:
lm(formula = WranglerSales ~ Year + MonthFactor + WranglerQueries +
    CPI.All + CPI.Energy, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
```

```
-2027.49  -509.49   -80.45   414.44  2291.31

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           4.192e+06  2.283e+06   1.837 0.071560 .
Year                 -2.151e+03  1.164e+03  -1.848 0.069871 .
MonthFactorAugust    -6.573e+02  6.748e+02  -0.974 0.334204
MonthFactorDecember  -1.511e+03  9.443e+02  -1.600 0.115282
MonthFactorFebruary  -2.280e+03  5.692e+02  -4.006 0.000184 ***
MonthFactorJanuary   -2.828e+03  6.236e+02  -4.535 3.08e-05 ***
MonthFactorJuly      -7.014e+02  6.635e+02  -1.057 0.295012
MonthFactorJune      -1.642e+02  6.039e+02  -0.272 0.786750
MonthFactorMarch     -7.035e+00  5.336e+02  -0.013 0.989528
MonthFactorMay        1.576e+03  5.466e+02   2.883 0.005574 **
MonthFactorNovember  -2.804e+03  8.574e+02  -3.271 0.001840 **
MonthFactorOctober   -1.611e+03  7.996e+02  -2.015 0.048672 *
MonthFactorSeptember -1.756e+03  7.048e+02  -2.492 0.015686 *
WranglerQueries       1.636e+02  3.267e+01   5.007 5.83e-06 ***
CPI.All               6.719e+02  2.962e+02   2.268 0.027194 *
CPI.Energy           -6.011e+01  2.851e+01  -2.108 0.039482 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 909.7 on 56 degrees of freedom
Multiple R-squared:  0.9527, Adjusted R-squared:  0.9401
F-statistic: 75.26 on 15 and 56 DF,  p-value: < 2.2e-16
```

d.) The final model I use to predict the sales of Wrangler include the following independent variables: “WranglerQueries”, “Year”, “CPI.All”, “CPI.Energy” and “MonthFactor”. I include these variables since the model with these variables produces a high R^2 value on the training data and all the variables have “*” or “.” notation associated with them.
When I apply the prediction model to the test data set, I obtain the following quality metrics (R output copied below):
OSR^2 =0.672
MAE= 2318.6
RMSE = 2631.8

The OSR^2 value (0.672) obtained on test data set is not as high as R^2 value (0.95) obtained on the training data set, which is not a good indicator for my model. However the OSR^2 value is still higher than 0.5 which means that the model still creates useful prediction for FCA rather than baseline model.

```
> SSE = sum((WranglerPredictions-test$WranglerSales)^2)
> SST = sum((test$WranglerSales-mean(train$WranglerSales))^2)
> R2=1-SSE/SST
> R2
[1] 0.6720774
> MAE=sum(abs(WranglerPredictions-test$WranglerSales))/12
> MAE
[1] 2318.643
> RMSE=sqrt(sum((WranglerPredictions-test$WranglerSales)^2)/12)
> RMSE
[1] 2631.917
```

e.) For Elantra sales, I included the following independent variables: “ElantraQueries” and “CPI.All” index. I removed the variables “Year”, “CPI.Energy” which were part of my model in Part-a.

In this model, I observe that all my independent variables have p-values lower than 0.05 which is good. However I observe that my R-squared value is 0.4941 which is quite lower than the R^2 value of the model I created for Wrangler sales in part-a. Also when I apply my model to the test data of 2016, I calculate a negative OSR^2 value of -1.029 showing that my model is not doing a good job in predicting the 2016s sales for Elantra.

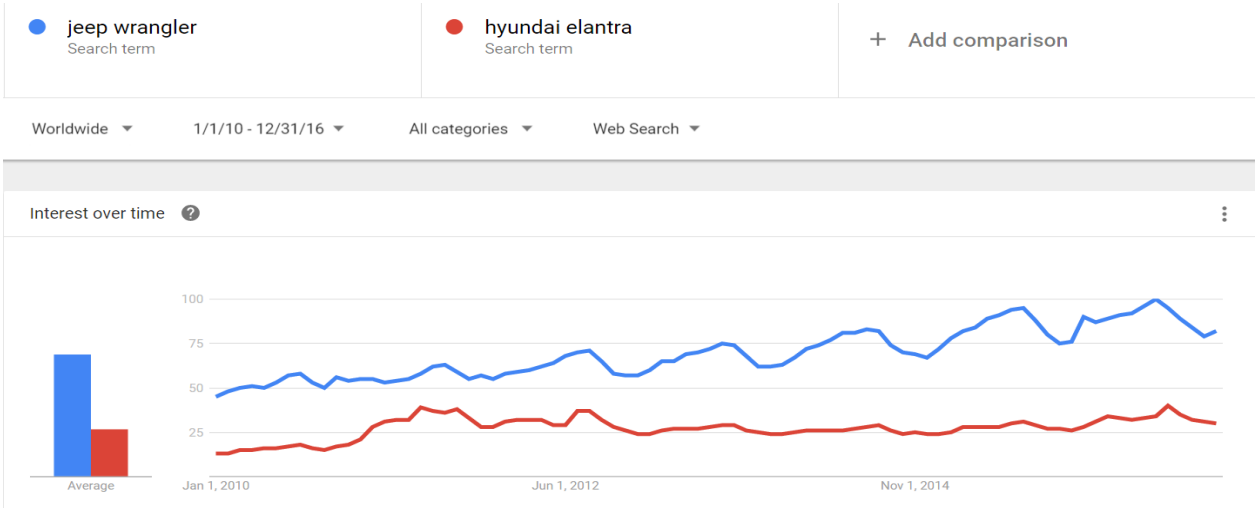
Therefore I expect the model found in Part-a. for Wrangler sales to have better predictive performance in 2017 due to its higher R^2 value.

R outputs:

```
> model25 = lm(ElantraSales ~ ElantraQueries + CPI.All, data=train)
> summary(model25)

Call:
lm(formula = ElantraSales ~ ElantraQueries + CPI.All, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-5965.8 -2552.5  -242.5  2157.1  8851.6

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    -66699.94   14783.93  -4.512 2.57e-05 ***
ElantraQueries    480.32     112.16   4.283 5.86e-05 ***
CPI.All           326.62      67.11   4.867 6.90e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3601 on 69 degrees of freedom
Multiple R-squared:  0.4941, Adjusted R-squared:  0.4794
F-statistic: 33.69 on 2 and 69 DF,  p-value: 6.179e-11
```

```
> SSE = sum((ElantraPredictions-test$ElantraSales)^2)
> SST = sum((test$ElantraSales-mean(train$ElantraSales))^2)
> R2=1-SSE/SST
> SSE
[1] 395403231
> SST
[1] 194788582
> R2
[1] -1.02991
```

f.) Google Trends Comparison:



I observe that Jeep Wrangler search trend displays growth over years and it also displays a clear seasonality however there is no clear seasonality in Hyundai Elantra search trend. So I do not think including seasonality would be beneficial to the linear model to predict the Elantra sales.

g.) Various costs associated producing more vehicles than demanded include:

- inventory holding cost
- depreciation cost of the vehicle for 1 month
- opportunity cost of producing another car model in your factory that has more demand in April.

Various costs associated producing less than demand include:

- cost of losing a sales/profit opportunity that month if the customer buys a different brand car
- loss of interest on the money you make if you sell the car in the April instead of the next month
- customer dissatisfaction due to not meeting customer demand

The answer to this production level question depends on the Underage Cost and the Overage Cost for the product. Underage cost is the financial loss you will suffer due to not satisfying one more unit demanded and Overage cost is the financial loss you will suffer due to having one more unsold vehicle in your inventory. A newsvendor model (which is a supply chain model) can be deployed to compute the optimal number of units to produce in March to meet the demand of April.

If the Underage (Understocking) cost is relatively high compared to the Overage Cost, then producing fewer than the mean demand (18081) is a better idea.
If the Overage (Overstocking) cost is relatively high compared to the Underage Cost , then producing more than the mean demand (18081) is a better idea.
Therefore when we develop models for management decisions, we should not only focus on predicting the demand but also the Expected Values of gain and loss we will face when our prediction does not meet the actual demand.

APPENDIX – R Code

R Script:

```
wranglerelantra = read.csv("wranglerElantra.csv")

str(wranglerelantra)

train = subset (wranglerelantra, Year <= 2015)

test = subset (wranglerelantra, Year > 2015)

str(train)

str(test)

model1 = lm(WranglerSales ~ Year + Unemployment + WranglerQueries + CPI.All +
CPI.Energy, data=train)

summary(model1)

model2 = lm(WranglerSales ~ Unemployment + WranglerQueries + CPI.All + CPI.Energy,
data=train)

summary(model2)

Year = c(wranglerelantra$Year)

WranglerSales = c(wranglerelantra$WranglerSales)

Unemployment = c(wranglerelantra$Unemployment)

WranglerQueries=c(wranglerelantra$WranglerQueries)

CPI.All = c(wranglerelantra$CPI.All)

CPI.Energy = c(wranglerelantra$CPI.Energy)

onedata=data.frame(WranglerSales,Year,Unemployment,WranglerQueries,CPI.All,
CPI.Energy)

cor(onedata[,c("WranglerSales", "Year", "Unemployment", "WranglerQueries","CPI.All",
"CPI.Energy")])

model3 = lm(WranglerSales ~ Year  + WranglerQueries + CPI.All + CPI.Energy,
data=train)

summary(model3)

model4 = lm(WranglerSales ~ Year + WranglerQueries + CPI.All, data=train)

summary(model4)

model5 = lm(WranglerSales ~ WranglerQueries + CPI.All + CPI.Energy, data=train)

summary(model5)

model6 = lm(WranglerSales ~ Year + WranglerQueries + CPI.All, data=train)

summary(model6)

model7 = lm(WranglerSales ~ WranglerQueries + CPI.All + CPI.Energy, data=train)

summary(model7)
```

```
model8 = lm(WranglerSales ~ Year  + WranglerQueries + CPI.Energy, data=train)

summary(model8)

model9 = lm(WranglerSales ~ WranglerQueries + CPI.All + CPI.Energy, data=train)

summary(model9)

model1 = lm(WranglerSales ~ Year + Unemployment + WranglerQueries + CPI.All +
CPI.Energy, data=train)

summary(model1)

model10 = lm(WranglerSales ~ Unemployment + WranglerQueries + CPI.Energy, data=train)

summary(model10)

summary(model3)

model11 = lm(WranglerSales ~ Year + MonthFactor + Unemployment + WranglerQueries +
CPI.All + CPI.Energy, data=train)

summary(model11)

model12 = lm(WranglerSales ~ Year + MonthFactor + WranglerQueries + CPI.All +
CPI.Energy, data=train)

summary(model12)

model13 = lm(WranglerSales ~ Year + MonthFactor + WranglerQueries + CPI.All +
CPI.Energy, data=train)

model21 = lm(ElantraSales ~ Year + Unemployment + ElantraQueries + CPI.All +
CPI.Energy, data=train)

summary(model21)

model22 = lm(ElantraSales ~ Year + ElantraQueries + CPI.All + CPI.Energy, data=train)

summary(model22)

model22 = lm(ElantraSales ~ Year + Unemployment + ElantraQueries + CPI.All,
data=train)

summary(model22)

model23 = lm(ElantraSales ~ Year + ElantraQueries + CPI.All, data=train)

summary(model23)

model24 = lm(ElantraSales ~ Year + ElantraQueries, data=train)

summary(model24)

model25 = lm(ElantraSales ~ ElantraQueries + CPI.All, data=train)

summary(model25)

ElantraPredictions = predict(model25, newdata=test)

summary(ElantraPredictions)

str(ElantraPredictions)
```

```
SSE = sum((ElantraPredictions-test$ElantraSales)^2)

SSE

SST = sum((test$ElantraSales-mean(train$ElantraSales))^2)

SST

R2=1-SSE/SST

R2

train1 = subset (wranglerelantra, MonthFactor == "January" | MonthFactor == "February"
| MonthFactor == "May" | MonthFactor == "September" | MonthFactor == "October" |
MonthFactor == "November", Year <= 2015)

train1

model12

model121 = lm(WranglerSales ~ Year + MonthFactor + WranglerQueries + CPI.All +
CPI.Energy, data=train1)

summary(model121)


WranglerPredictions = predict(model12, newdata=test)

summary(WranglerPredictions)

str(WranglerPredictions)

SSE = sum((WranglerPredictions-test$WranglerSales)^2)

SSE

SST = sum((test$WranglerSales-mean(train$WranglerSales))^2)

SST

R2=1-SSE/SST

R2

MAE=sum(abs(WranglerPredictions-test$WranglerSales))

MAE

WranglerPredictions-test$WranglerSales

MAE=sum(abs(WranglerPredictions-test$WranglerSales))/12

MAE

RMSE=sqrt(sum((WranglerPredictions-test$WranglerSales)^2)/12)

RMSE
```