```
> # DARIEN OLIVA - ECON 453 - FINAL EXAM PART 2
>
> # IMPORT DATA FOR ALL QUESTIONS/SHEETS
>
> library(readxl)
> data1 <- read_excel("Downloads/final_exam_part2_data_version_a.xlsx",
+                     sheet = "Housing_Starts")
> View(data1)
>
> library(readxl)
> data2 <- read_excel("Downloads/final_exam_part2_data_version_a.xlsx",
+                     sheet = "gym")
> View(data2)
>
> library(readxl)
> data3 <- read_excel("Downloads/final_exam_part2_data_version_a.xlsx",
+                     sheet = "Houses")
> View(data3)
>
> # QUESTION 1 A
>
> model1 <- lm(data1$HStarts ~ data1$Date, data1)
>
> summary(model1)

Call:
lm(formula = data1$HStarts ~ data1$Date, data = data1)

Residuals:
     Min       1Q   Median       3Q      Max
-25.9665  -8.2120   0.1108   8.7541  24.3740

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.649e+02  2.318e+01  -11.43   <2e-16 ***
data1$Date   2.462e-07  1.634e-08   15.07   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.48 on 93 degrees of freedom
Multiple R-squared:  0.7095,  Adjusted R-squared:  0.7063
F-statistic: 227.1 on 1 and 93 DF,  p-value: < 2.2e-16

>
```

```
> plot(data1$Date,data1$HStarts, type = "l", main = "Linear Trend Model For Housing Starts",
+     xlab = "Date",
+     ylab = "Housing Starts")
> lines(data1$Date, predict(model1),
+      col = 2,
+      lwd = 2)
>
> # The p-value is 2.2e-16 which indicates that the relationship between Date and Housings
Starts
> # is highly statistically significant.
>
> ts(data1$HStarts)
Time Series:
Start = 1
End = 95
Frequency = 1
 [1]  40.2  35.4  49.9  49.0  54.0  60.5  57.6  54.5  58.8  53.2  53.0  42.7  47.2
[14]  49.7  58.0  66.8  67.8  74.7  69.2  69.0  75.8  77.0  62.2  63.2  58.7  66.1
[27]  83.3  76.3  87.2  80.7  84.0  80.4  78.4  78.4  83.8  67.6  60.7  65.1  80.2
[40]  94.9  92.5  87.3 101.0  86.2  94.2  92.0  75.8  73.4  73.0  61.9  79.7 108.5
[53]  99.6 112.3 107.2  99.2 111.6  90.9  89.9  78.1  74.3  84.1  90.7 106.2 105.0
[66] 111.6 115.2 102.8  95.0 114.5  87.8  86.5  82.3  87.8  97.1 105.2 106.0 116.3
[79] 112.3 102.6 104.4 109.6  97.9  81.4  91.6  89.7 107.2 117.5 123.7 112.0 111.9
[92] 113.8 109.7 106.1  95.9
> housingstarts <- ts(data1$HStarts, frequency = 12, start = 2011)
> autoplot(housingstarts, main = " Housings Starts for 2011-2018", xlab = "Date", ylab =
"Housing Starts")
>
> autoplot(housingstarts, main = "Forecast for Housing Starts",xlab = "Date", ylab = "Housing
Starts")+autolayer(hw(housingstarts, seasonal = "multiplicative", PI=FALSE))
+autolayer(hw(housingstarts, seasonal = "additive", PI=FALSE), col="Red")
>
>
> # QUESTION 1 B
>
> data1 <- data1 %>%
+   mutate(Month=month(ymd(Date)))
> View(data1)
>
> data1 <- data1 %>%
+   mutate(Month = format(as.Date(Date), "%m"))
>
> data1$Jan=ifelse((data1$Month=="01"),1,0)
> data1$Feb=ifelse((data1$Month=="02"),1,0)
```

```
> data1$Mar=ifelse((data1$Month=="03"),1,0)
> data1$Apr=ifelse((data1$Month=="04"),1,0)
> data1$May=ifelse((data1$Month=="05"),1,0)
> data1$June=ifelse((data1$Month=="06"),1,0)
> data1$July=ifelse((data1$Month=="07"),1,0)
> data1$Aug=ifelse((data1$Month=="08"),1,0)
> data1$Sep=ifelse((data1$Month=="09"),1,0)
> data1$Oct=ifelse((data1$Month=="10"),1,0)
> data1$Nov=ifelse((data1$Month=="11"),1,0)
> data1$Dec=ifelse((data1$Month=="12"),1,0)
>
> model1b <- lm(HStarts ~ Feb + Mar + Apr + May + June + July + Aug + Sep + Oct + Nov +
Dec, data = data1)
> summary(model1b)

Call:
lm(formula = HStarts ~ Feb + Mar + Apr + May + June + July +
   Aug + Sep + Oct + Nov + Dec, data = data1)

Residuals:
   Min      1Q  Median      3Q     Max
-41.550 -13.469   3.012  15.987  31.725

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   66.000      6.991   9.441 8.58e-15 ***
Feb            1.475      9.887   0.149  0.88177
Mar           14.762      9.887   1.493  0.13919
Apr           24.550      9.887   2.483  0.01504 *
May           25.975      9.887   2.627  0.01025 *
June          28.425      9.887   2.875  0.00513 **
July          28.800      9.887   2.913  0.00460 **
Aug           22.563      9.887   2.282  0.02504 *
Sep           24.988      9.887   2.527  0.01339 *
Oct           24.213      9.887   2.449  0.01643 *
Nov           14.788      9.887   1.496  0.13853
Dec            4.414     10.234   0.431  0.66733
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.77 on 83 degrees of freedom
Multiple R-squared:  0.2311,   Adjusted R-squared:  0.1292
F-statistic: 2.268 on 11 and 83 DF,  p-value: 0.01792
```

```
>
>
> # QUESTION 1 C
>
> # SEE SCANNED ANSWERS
> # COMPARE THE MODELS
>
>
> # QUESTION 2 A
>
> # make dummy variables and make model 2
>
> data2$D_Plan=ifelse((data2$Plan=="Single"),1,0)
> summary(data2)
    Renewed          Age            Income           Plan              D_Plan
 Min.   :0.000   Min.   :20.00   Min.   : 30.00   Length:200         Min.   :0.000
 1st Qu.:0.000   1st Qu.:34.00   1st Qu.: 56.00   Class :character   1st Qu.:0.000
 Median :1.000   Median :50.00   Median : 96.50   Mode  :character   Median :1.000
 Mean   :0.575   Mean   :49.17   Mean   : 91.85                      Mean   :0.535
 3rd Qu.:1.000   3rd Qu.:65.00   3rd Qu.:122.00                      3rd Qu.:1.000
 Max.   :1.000   Max.   :80.00   Max.   :150.00                      Max.   :1.000
>
> model2 <- lm(Renewed~Age+Income+D_Plan, data = data2)
> summary(model2)

Call:
lm(formula = Renewed ~ Age + Income + D_Plan, data = data2)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.9699 -0.2519  0.0438  0.2910  0.9508 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.3380805  0.1107921  -3.051  0.00259 ** 
Age          0.0054229  0.0016099   3.368  0.00091 ***
Income       0.0077886  0.0007725  10.082  < 2e-16 ***
D_Plan      -0.1289184  0.0562271  -2.293  0.02292 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3952 on 196 degrees of freedom
Multiple R-squared:  0.3735,  Adjusted R-squared:  0.3639 
F-statistic: 38.95 on 3 and 196 DF,  p-value: < 2.2e-16
```

```
>
> # use this code to run the prediction for 40 year old man
>
> newdata <- data.frame(Age = 40, Income = 80, D_Plan = 1)
> predictions <- predict(model2, newdata = newdata)
> predictions
        1
0.3730033
>
> # 1
> # 0.3730033
>
> # QUESTION 2 B
>
>
> logistic_model <- glm(formula = Renewed ~ Age + Income + D_Plan, family = binomial, data =
data2)
> logistic_model

Call:  glm(formula = Renewed ~ Age + Income + D_Plan, family = binomial,
    data = data2)

Coefficients:
(Intercept)          Age       Income       D_Plan
   -4.92665      0.03699      0.04372     -0.85931

Degrees of Freedom: 199 Total (i.e. Null);  196 Residual
Null Deviance:     272.7
Residual Deviance: 185.7      AIC: 193.7
> predictions <- predict(logistic_model, newdata = data.frame(Age = 40, Income = 80, D_Plan =
1))
> predictions
         1
-0.8084314
> # 1
> # -0.8084314
>
> # RESULT : Classify as non-renew
>
> #calculate odds.
> odds <- predictions / (1 - predictions)
> odds
         1
```

```
-0.4470346
>  # 1
> # -0.4470346
>
>
> # QUESTION 2 C
>
> # COMPARE THE MODELS
>
> # QUESTION 3 A
>
> model3 <- lm(Price ~ Sqft + Beds + Baths + Colonial, data = data3)
> summary(model3)

Call:
lm(formula = Price ~ Sqft + Beds + Baths + Colonial, data = data3)

Residuals:
    Min      1Q  Median      3Q     Max
-157118  -47479   -4742   38849  168327

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 165888.66   52353.54   3.169  0.00343 **
Sqft            91.68      32.34   2.834  0.00801 **
Beds          4372.36   18561.84   0.236  0.81533
Baths        66619.61   24659.48   2.702  0.01109 *
Colonial     74557.88   27374.26   2.724  0.01051 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 68440 on 31 degrees of freedom
Multiple R-squared:  0.777,	Adjusted R-squared:  0.7483
F-statistic: 27.01 on 4 and 31 DF,  p-value: 1.031e-09

>
> # QUESTION 3 B
>
> # 91.68, meaning on average, for each additional sq ft in a home, the sales price increases
> # by $91.87 (all else constant). This is statistically significant as indicated by the low p-value.
>
> # QUESTION 3 C
>
```

```
> #74,557.88, meaning that colonial homes are expected to sell for $74,557.88 more than
non-colonial
> # homes on average, and while controlling other variables. This is statistically significant as
indicated by low p-value.
>
>
> # QUESTION 3 D
>
> # Variables Sqft, Bath, and Colonial are all statistically significant as indicated by their low
p-values.
> # This shows a strong relationship between these variables and sales price. Beds is not a
statistically significant variable
> # as indicated by its high p-value. This means there is not a strong relationship between
number of Bedrooms and sales price.
>
> # QUESTION 3 E
>
> # 95 % confidence interval for expected price for 2500 sqft, colonial, 3 bed, 2 bath.
>
> predict.lm(model3, data.frame(Sqft = 2500, Beds = 3, Baths = 2, Colonial = 1),
+            interval = "predict", level = 0.95)
      fit      lwr      upr
1 615990.9 466436.4 765545.4
>
> #RESULTS:
> # fit      lwr      upr
> # 615990.9 466436.4 765545.4
```