```
survey<-read.csv("E:/Repos/StatisticsR/DSC520-Statistics/week4/acs-14-1yr-s0201.csv")

head(survey)


library(readxl)


housing<-read_excel("E:/Repos/StatisticsR/DSC520-Statistics/week4/week-6-housing.xlsx")

head(housing)


# Use the apply function on a variable in your dataset


apply(survey,2,length)

apply(housing,2,length)


# Use the aggregate function on a variable in your dataset


aggregate(survey$Geography,list(unique.values=survey$Geography),length)




# Use the plyr function on a variable in your dataset - more specifically,

# I want to see you split some data, perform a modification to the data, and then bring it back together


d <- data.frame(year = rep(2000:2002, each = 3),count = round(runif(9, 0, 20)))

print(d)


library(plyr)

ddply(d, "year", function(x) {

  mean.count <- mean(x$count)

  sd.count <- sd(x$count)

  cv <- sd.count/mean.count
```

```
  data.frame(cv.count = cv)
})
ddply(d, "year", summarise, mean.count = mean(count))
ddply(d, "year", transform, total.count = sum(count))
ddply(d, "year", mutate, mu = mean(count), sigma = sd(count),cv =
sigma/mu)


housing.dat <- subset(housing, 'sale year' > 2000)
x <- ddply(housing.dat, c("'sale year'", "ctyname"), summarize, homeruns =
sum(housing.dat$`Sale Price`))
head(x)


# Check distributions of the data
install.packages("fitdistrplus")
library(fitdistrplus)
normal_dist <- fitdist(housing$`Sale Price`, "norm")
plot(normal_dist)


# Identify if there are any outliers
summary(housing$`Sale Price`)
hist(housing$`Sale Price`,xlab = "Price",main = "Histogram of
Price",breaks = sqrt(nrow(housing.dat)))


# Create at least 2 new variables
country<-rep("USA",12865)
serial_no<-c(1:12865)


new_df<-cbind(serial_no,housing,country)
head(new_df)



#Explain any transformations or modifications you made to the dataset
```

```
#Answer - Splitted the data on the basis of year ranging from 2000 to 2002
and used ddply and brought back to the dataset



#Create two variables; one that will contain the variables Sale Price and

#Square Foot of Lot (same variables used from previous assignment on
simple regression)

#and one that will contain Sale Price and several additional predictors of
your choice.

#Explain the basis for your additional predictor selections.


#Answer -

install.packages("olsrr")

library(olsrr)

new_model<-lm(new_df$`Sale Price`~new_df$bedrooms,data=new_df)

new_model1<-lm(new_df$`Sale Price`~new_df$sq_ft_lot,data=new_df)

model_final <- lm(new_df$`Sale Price`~
building_grade+square_feet_total_living+bedrooms +sq_ft_lot+
new_df$bath_full_count,data=new_df)

#This had the highest Adjusted R-Squared value


#Execute a summary() function on two variables defined in the previous
step to compare the model results.

#What are the R2 and Adjusted R2 statistics?

#Explain what these results tell you about the overall model.

#Did the inclusion of the additional predictors help explain any large
variations found in Sale Price?


#Answer-

summary(new_model1)

# R-square -0.01435

#Adjusted R-square - 0.01428

#This shows that the square feet have 1.4% of the variation in the sale
Price

summary(model_final)
```

```
# R-square -0.2166

#Adjusted R-square - 0.2163

#This shows that the all the factors have 21.6% of the variation in the
sale Price



#Considering the parameters of the multiple regression model you have
created.

#What are the standardized betas for each parameter and what do the values
indicate?


#Answer- value increases by one, SD sale price increases by that variables
beta value assuming the other are held constant



#Calculate the confidence intervals for the parameters in your model and
explain what the results indicate.


#Answer -

l.model <- lm(new_df$`Sale Price` ~ 1, new_df)

confint(l.model, level=0.95)

#So the population of Sale price between 2.5% to 97.5% is 653749.4 and
667726.1


#Assess the improvement of the new model compared to your original model
(simple regression model)

#by testing whether this change is significant by performing an analysis
of variance.


#Answer-

library(ggplot2)

ggplot(new_df,aes(`Sale Price`,sq_ft_lot))+ geom_point()
+stat_smooth(method = lm)
```

```
#the true value of the regression coefficients will lie within the bounds
of the confidence interval



#Perform casewise diagnostics to identify outliers and/or influential
cases,

#storing each function's output in a dataframe assigned to a unique
variable name.


#Answer -

anova(new_model,final_model)



#Calculate the standardized residuals using the appropriate command,
specifying those that are +-2,

#storing the results of large residuals in a variable you create.


#Answer-

model <- lm(formula = `Sale Price` ~ sq_ft_lot + square_feet_total_living

            +bedrooms, data = new_df)

summary(model)


res <- residuals(model)


res <- as.data.frame(res)

# plot into histogram

ggplot(res, aes(res)) + geom_histogram(alpha = 0.5,bins=70) +
labs(title="Plotting Residuals")


plot(model)


#Use the appropriate function to show the sum of large residuals.


#Answer -
```

```
res <- as.data.frame(res)


#Which specific variables have large residuals (only cases that evaluate
as TRUE)?


#Answer - As per the result Sale price has the large residuals



#Investigate further by calculating the leverage, cooks distance, and
covariance rations.

#Comment on all cases that are problematics.


#Answer-

plot(hatvalues(l.model), pch=23, bg='orange', cex=2, ylab='Hat values')

plot(cooks.distance(l.model), pch=23, bg='orange', cex=2, ylab="Cook's
distance")




#Perform the necessary calculations to assess the assumption of no
multicollinearity and

#state if the condition is met or not.


#Answer-

install.packages("VIF")

library(VIF)

car::vif(final_model)



#Visually check the assumptions related to the residuals using the plot()
and hist() functions.

#Summarize what each graph is informing you of and if any anomalies are
present.
```

#Answer -

plot(model)

#Show that the residual and fitted value increases in direct proportion.

#Overall, is this regression model unbiased?

#If an unbiased regression model, what does this tell us about the sample vs. the entire population model?

#Answer - Yes the model is unbiased. It saying that the sample is similar to the entire population model/