HW 1

library(tidyverse) library(ggplot2) library(rsample) library(caret) library(modelr) library(parallel) library(foreach)

gas_prices = read.csv('/Users/franklinstudent/Desktop/GitHub/ECO395M/data/GasPrices.csv')

1A

ggplot(data=gas_prices) + geom_boxplot(aes(x=Competitors, y=Price, fill = Competitors))

In this boxplot, we are assesing whether the lack of direct competition in sight results in higher gas prices. The boxplot is separated into twocategories, yes or no, on whether a competitor is within sight of the observed gas station. I also added color to better differentiate between both categories. When interpreting the boxplot, it becomes clear that the lack of direct competition in sight results in higher gas prices. The median price of gas stations with no competitors in sight is approximately $1.89/gal, while in contrast, the median price of gas stations with competitors in sight is $1.85/gal. Moreover, gas stations with no direct competitors in sight have a larger range of prices, relative to gas stations with direct competitors in sight.

1B

ggplot(data = gas_prices) + geom_point(mapping = aes(x = Income, y = Price))

In this scatterplot, we are determining whether more affluent areas experience higher gas prices. Interpreting the data, there appears to be a positive association between the variables Income and Price, indicating yes, more affluent areas experience higher gas prices. Although, it should be noted that there are several outliers.

1C

ggplot(data = gas_prices) + geom_col(mapping = aes(x = Brand, y = Price, fill = Brand))

In creating this barplot, we are determining if Shell charges more for gasoline relative to other brands. Upon observing the barplot, it becomes clear that Shell charges more for gas than either Chevron-Texaco or ExxonMobil, but charges less than Other.

1D

ggplot(data = gas_prices) + geom_histogram(aes(x = Price, after_stat(density)), binwidth = 0.05) + facet_wrap(~Stoplight)

In this set of data, we created a faceted histogram to assist in determing if whether the precence of stoplights near gas stations increases gas prices. Observing the histograms, gas stations at stoplights have a higher median gas price relative to gas stations without stoplights. The median price of gasoline at gas stations without a stoplight is $1.80/gal, while the median price of gasoline at gas stations wiht a stoplight is $1.90/gal. Additionally, it should be noted that there is an outlier of $2.10/gal for gas stations without a stoplight.

1E

```
ggplot(data=gas_prices) + geom_boxplot(aes(x=Highway, y=Price, fill = Highway))
```

I created boxplots to assist in determining if a gas station has higher gas prices as a consequence of direct highway access. Upon observing the boxplots, it becomes clear that gas stations with direct highway access increases have higher gas prices. Gas stations with direct highway access have a higher median price of approximately $1.89/gal, while in contrast, gas statiosn without a direct highway access have a lower median price of approximately $1.84/gal.

2

```
rides = read.csv('/Users/franklinstudent/Desktop/GitHub/ECO395M/data/bikeshare.csv')
```

Plot A

```
ride_total = rides %>% group_by(hr) %>% summarize(ride_totals = sum(total))
```

```
ggplot(ride_total) + geom_line(aes(x = hr, y = ride_totals))
```

This line graph displays the change in average ridership throughout the day. The time of day lays on the x-axis, and the ridership totals lies on the y-axis. The graph shows that the busiest time of day for ridership is during peak rushhour.

Plot B

```
working_day = rides %>% filter(workingday == '1')
```

```
non_working_day = rides %>% filter(workingday == '0')
```

```
working_day_total = working_day %>% group_by(hr) %>% summarise(ride_ = sum(total))
```

```
non_working_day_total = non_working_day %>% group_by(hr) %>% summarise(ride_ = sum(total))
```

```
day_list = c('1', '0')
```

```
combined_rides = rides %>% filter(workingday %in% day_list) %>% group_by(hr, workingday) %>% summarize(total_rides = sum(total))
```

```
ggplot(combined_rides) + geom_line(aes(x = hr, y = total_rides, color = workingday)) + facet_wrap(~workingday)
```

This line faceted line graph was made to help determine the differences in ridership between workdays, which have a value of 1, and non-workdays, which have a value of 0. Upon observing either graph, its easy to understand that the highest peaks of ridership occurs during a workday, and at rushhour times. Non-workdays have a peak during mid-afternoons.

Plot C

```
rides = rides %>% filter(hr == '8') %>% mutate(day = ifelse(workingday == 1, 'Work Day', "Non-Work Day"))
```

```
d1 = rides %>% group_by(day, weathersit) %>% summarize(average_rides = mean(total))
```

```
ggplot(data = d1) + geom_col(mapping = aes(x = weathersit, y = average_rides, fill = weathersit), position = ‘dodge’) + facet_wrap(~day)
```

In addition to this faceted bar graph breaking down between non-workdays and workdays, differences of average ridership can also observed in reference to weather conditions. At first glace, its very clear that average ridership is much higher during a workday. In all cases, ridership decreases as the weather worsens.

3

```
abia = read.csv(‘/Users/franklinstudent/Desktop/GitHub/ECO395M/data/ABIA.csv’)
```

```
airlines = c(‘AA’, ‘UA’, ‘WN’, ‘CO’)
```

```
combined_airlines = abia %>% filter(UniqueCarrier %in% airlines) %>% group_by(UniqueCarrier, DayOfWeek) %>% summarise(total_count = n(), cancelled = sum(Cancelled == 1), delayed_percentage = cancelled/total_count)
```

```
ggplot(combined_airlines) + geom_line(aes(x = DayOfWeek, y = total_count, color = UniqueCarrier)) + scale_x_continuous(breaks = 1:7) + scale_y_log10()
```

I chose 4 airlines to observe and compare throughout the week. AA (American Airlines), CO (Continental Airlines), UA (United Airlines), and WN (Southwest Airlines). WN has the most flights at ARIA, while UA has, by far, the lowest number of flights. All airlines have a dip in traffic on Saturdays, before increasing again for Sunday. This makes intutive sense because people tend to fly someplace for the weekend, and return on Sunday for work on Monday.

```
ggplot(data = combined_airlines) + geom_col(mapping = aes(x = UniqueCarrier, y = delayed_percentage, fill = UniqueCarrier), position = ‘dodge’) + facet_wrap(~DayOfWeek)
```

I then decided to create a bar graph to display the percentages of canceled flights each day of the week for the the 4 airlines listed. One conclusion made is that although WN has the most flights out of ABIA, AA has the highest percentages of cancelled flights on nearly every day of the week, with Tuesday being the worst day to fly on AA, but the best day to fly UA.

4

```
sclass = read.csv(‘/Users/franklinstudent/Desktop/GitHub/ECO395M/data/sclass.csv’)
```

```
class350 = sclass %>% filter(trim == ‘350’)
```

```
class63 = sclass %>% filter(trim == ‘63 AMG’)
```

```
ggplot(data = class350) + geom_point(mapping = aes(x = mileage, y = price), color = ‘darkgrey’)
```

```
ggplot(data = class63) + geom_point(mapping = aes(x = mileage, y = price), color = ‘darkgrey’)
```

```
class350_split = initial_split(class350, prop = 0.8) class350_train = training(class350_split) class350_test = testing(class350_split)
```

```
class63_split = initial_split(class63, prop = 0.8) class63_train = training(class63_split) class63_test = testing(class63_split)
```

```
lm1 = lm(price ~ mileage, data = class350_train) lm2 = lm(price ~ poly(mileage, 2), data = class350_train)
```

```
lm3 = lm(price ~ mileage, data = class63_train) lm4 = lm(price ~ poly(mileage, 2), data = class63_train)
```

```
knn350 = knnreg(price ~ mileage, data = class350_train, k = 30) rmse(knn350, class350_test)
```

```
knn63 = knnreg(price ~ mileage, data = class63_train, k = 150) rmse(knn63, class63_test)
```

```
class350_test = class350_test %>% mutate(Price_pred = predict(knn350, class350_test))

class63_test = class63_test %>% mutate(Price_pred = predict(knn63, class63_test))

p_test350 = ggplot(data = class350_test) + geom_point(mapping = aes(x = mileage, y = price), alpha=0.2) p_test350

p_test63 = ggplot(data = class63_test) + geom_point(mapping = aes(x = mileage, y = price), alpha=0.2) p_test63

p_test350 + geom_line(aes(x = mileage, y = Price_pred), color = ‘red’, size = 1.5)

p_test63 + geom_line(aes(x = mileage, y = Price_pred), color = ‘red’, size = 1.5)

rmse_out350 = foreach(i=1:10, .combine=‘c’) %do% { class350_split = initial_split(class350, prop = 0.8) class350_train = training(class350_split) class350_test = testing(class350_split)

knn_model = knnreg(price ~ mileage, data = class350_train, k = 25) modelr::rmse(knn_model350, class350_test) }

rmse_out350

rmse_out63 = foreach(i=1:10, .combine=‘c’) %do% { class63_split = initial_split(class63, prop = 0.8) class63_train = training(class63_split) class63_test = testing(class63_split)

knn_model63 = knnreg(price ~ mileage, data = class63_train, k = 25) modelr::rmse(knn_model63, class63_test) }

rmse_out63
```

I believe that the trim 63 AMG has a higher optimal value of k because theres a larger concentration of cars valued much higher with mileage close to zero,whereas the 350 trim is much more spread out.