Project part 2
CODE:

```
library(ggplot2)
library(lubridate)
library(dplyr)
library(psych)
df = read.csv("C:\\Users\\dingj\\Downloads\\flights_sample_3m.csv")
df$FL_DATE <- as.Date(df$FL_DATE, format="%Y-%m-%d") #Date tim
df$month <- month(df$FL_DATE)
df_2019 <- subset(df, format(FL_DATE, "%Y") == "2019") #Use 2019 data
df_2019_sorted <- df_2019[order(df_2019$FL_DATE), ] #sort it by DATE
row.names(df_2019_sorted) <- NULL #reset index
df_2019_sorted = df_2019_sorted[sample(nrow(df_2019_sorted), 100000), ] #pick 100,000
sample
describe(df_2019_sorted$DISTANCE)
describe(df_2019_sorted$AIR_TIME)
df_2019_sorted$VELOCITY = df_2019_sorted$DISTANCE/df_2019_sorted$AIR_TIME
describe(df_2019_sorted$VELOCITY)

ggplot(data = df_2019_sorted, aes(x = DISTANCE, y = AIR_TIME)) +
  geom_point(alpha = 0.3, size = 1.5) +  # Adjust alpha for transparency and size for point
size
  geom_smooth(method = "lm", color = "red") +  # Adds a linear regression line
  theme_minimal() +
  labs(title = "Distance vs. Air Time",
       x = "Distance (Miles)",
       y = "Air Time (minutes)") +
  scale_color_gradient(low = "blue", high = "red")  # Color gradient from low to high density

# Calculate the average distance by month
average_distance_by_month <- df_2019_sorted %>%
  group_by(month) %>%
  summarise(Average_DISTANCE = mean(DISTANCE, na.rm = TRUE)) %>%
  ungroup()

# Create a bar chart to display the average distance by month
ggplot(average_distance_by_month, aes(x = as.factor(month), y = Average_DISTANCE)) +
  geom_bar(stat = "identity", fill = "dodgerblue3") +
  theme_minimal() +
  labs(title = "Average Distance by Month",
       x = "Month",
       y = "Average Distance (Miles)") +
```

```
  scale_x_discrete(name = "Month", labels = c('Jan', 'Feb', 'Mar', 'Apr', 'May', 'Jun', 'Jul', 'Aug',
'Sep', 'Oct', 'Nov', 'Dec'))  # Label months appropriately

# Calculate the average air time by month
average_air_time_by_month <- df_2019_sorted %>%
  group_by(month) %>%
  summarise(Average_AIR_TIME = mean(AIR_TIME, na.rm = TRUE)) %>%
  ungroup()  # Ungroup for plotting

# Create a bar chart to display the average air time by month
ggplot(average_air_time_by_month, aes(x = as.factor(month), y = Average_AIR_TIME)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  theme_minimal() +
  labs(title = "Average Air Time by Month",
       x = "Month",
       y = "Average Air Time (Minutes)") +
  scale_x_discrete(name = "Month", labels = c('Jan', 'Feb', 'Mar', 'Apr', 'May', 'Jun', 'Jul', 'Aug',
'Sep', 'Oct', 'Nov', 'Dec'))  # Label months appropriately

air_time_model <- lm(AIR_TIME ~ DISTANCE, data = df_2019_sorted)
model_summary <- summary(air_time_model)
print(model_summary)
confint(air_time_model, level = 0.95) #95% CI
plot(air_time_model$fitted.values, air_time_model$residuals)
abline(h = 0, col = 'red')
title('Residuals vs Fitted')
qqnorm(air_time_model$residuals)
qqline(air_time_model$residuals, col = 'red')

new_data <- data.frame(DISTANCE = 5000)
predicted_air_time <- predict(air_time_model, newdata = new_data)
predicted_air_time #602.6423
predicted_air_time_interval <- predict(air_time_model, newdata = new_data, interval =
"predict")
predicted_air_time_interval #(577.3874,627.8971)

df_2019_sorted$DISTANCE
rows_with_long_distance <- subset(df_2019_sorted, DISTANCE > 5000, AIR_TIME)
#Only 3 flights with 5095 distance
```

Answer:

By chronological order in bullet point fashion:

- Import data
- Make sure the column 'FL_DATE' is in date time format and got month column
- Set year to 2019, and sample to 100,000
- Describe() the data for descriptive analysis
- Created a new column called Velocity by (DISTANCE/AIR_TIME)
- Graphed (scatter plot) Distance vs Airtime to check for Linear Regression
- Calculate average distance/air time per month for graphs that depicts average distance/air time per month.
- Graphed them.
- Made airtimemodel with lm() function
- Printed the summary of airtimemodel to gain inferential statistic data
- Plotted Residual Fitted Graph and QQPlot to check for normality, linearity, and homoscedasticity
- Created new data frame for prediction
- I set the distance to 5000 for prediction
- Predicted airtime = 602.6423
- Predicted airtime interval
- Found the actual data of flights in 2019 with distance >5000 and their airtime through data wrangling
- Compare the predicted airtime with the average of those 3's airtime
-