Plan-372 HW4 Write-Up

2024-11-05

```
knitr::opts_chunk$set(echo = FALSE)
```

Link to GitHib Repository: https://github.com/annamitchell23/plan372-hw4

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x readr::col_factor() masks scales::col_factor()
## x purrr::discard()    masks scales::discard()
## x dplyr::filter()     masks stats::filter()
## x dplyr::lag()        masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

Question 1:

```
## # A tibble: 5,022 x 3
##    origin dest  passengers
##    <chr>  <chr>      <dbl>
##  1 EWR    DTW       205100
##  2 FLL    DTW       378280
##  3 JAX    DTW        68740
##  4 LAX    DTW       422730
##  5 LAX    OAK       418340
##  6 LAX    PHX       662760
##  7 LGA    DTW       382450
##  8 LGA    MIA       750650
##  9 MCO    DTW       576280
## 10 MIA    DTW       204770
## # i 5,012 more rows
```

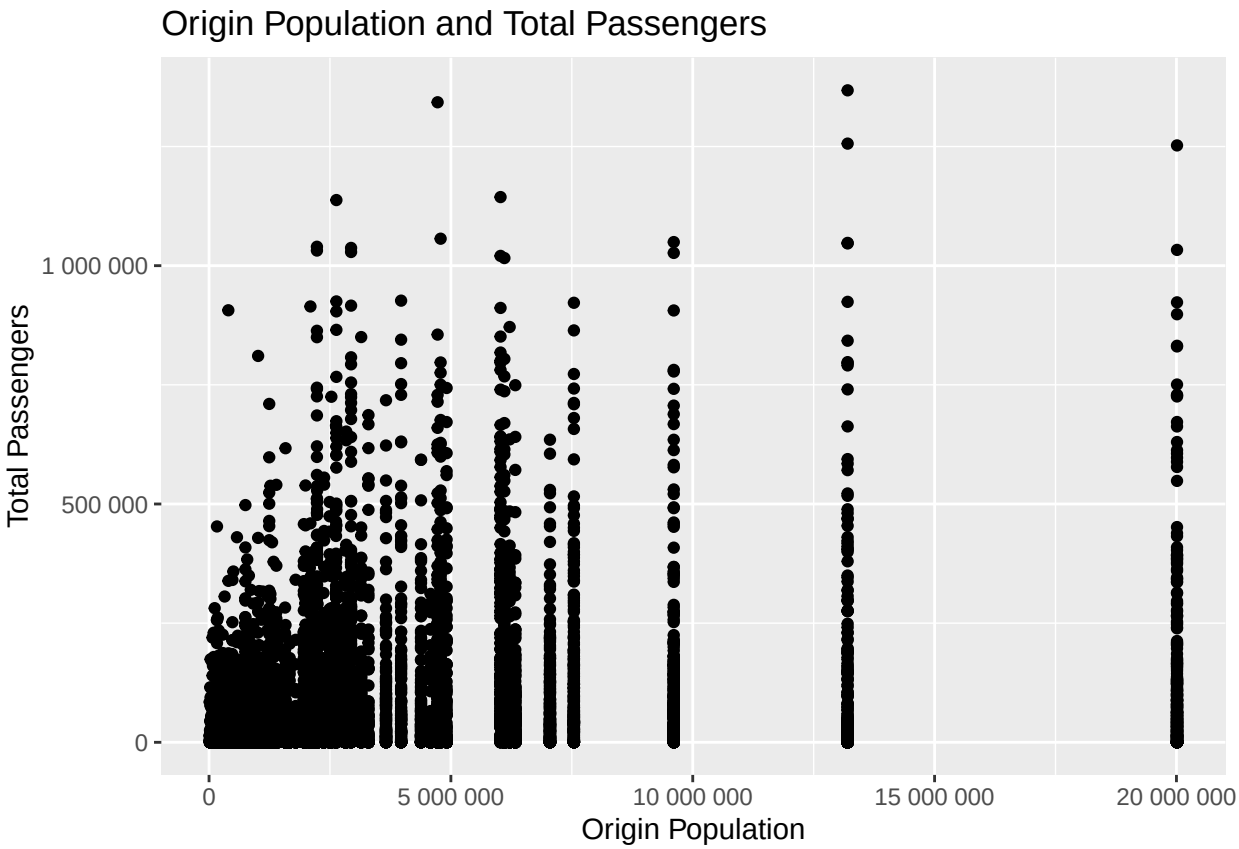
Question 2:

```
## To install your API key for use in future sessions, run this function with `install = TRUE`.
```

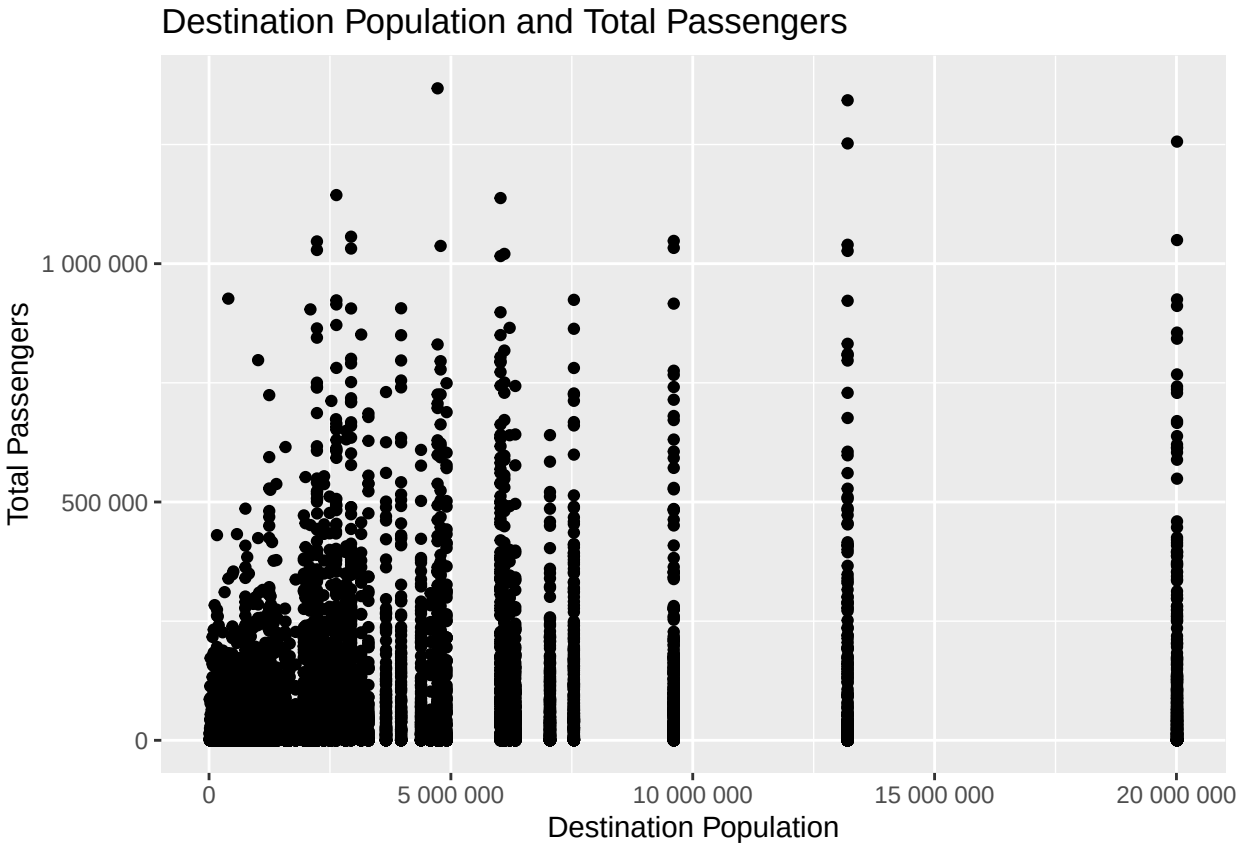
```
## Getting data from the 2017-2021 5-year ACS
```

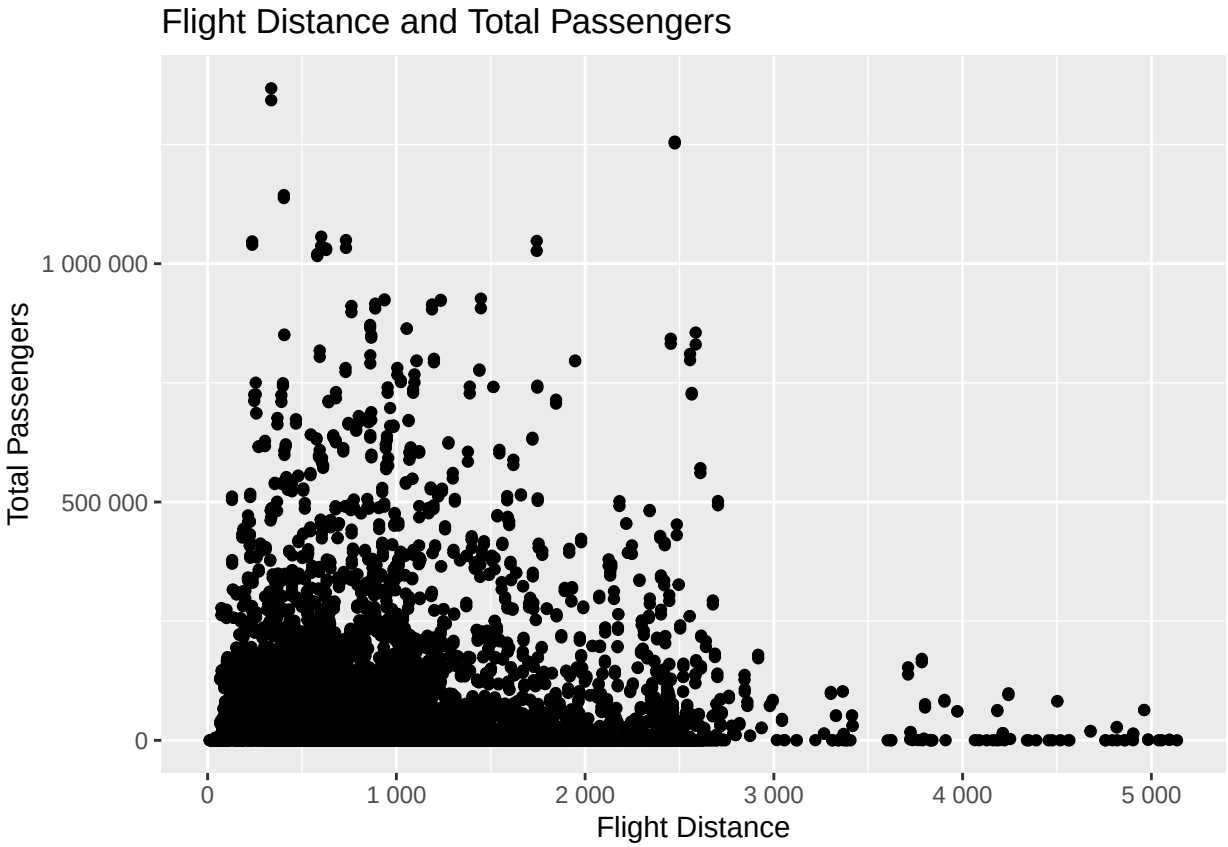
```
## `summarise()` has grouped output by 'origin_cbsa', 'dest_cbsa',
## 'origin_population', 'dest_population', 'origin_median_income',
## 'dest_median_income', 'origin_per_capita_income', 'dest_per_capita_income',
## 'origin_median_home_value', 'dest_median_home_value'. You can override using
## the `.groups` argument.
```

```
## Warning: Removed 98 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

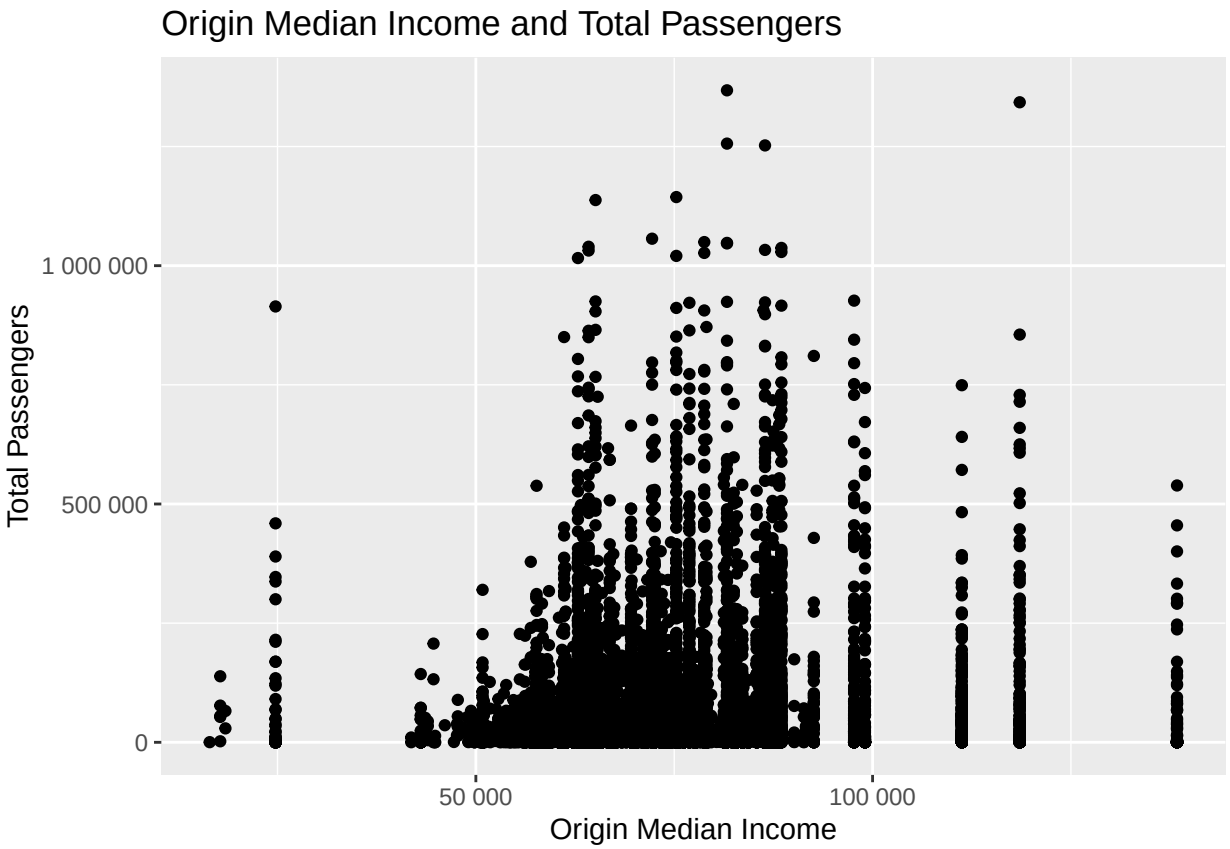



```
## Warning: Removed 99 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

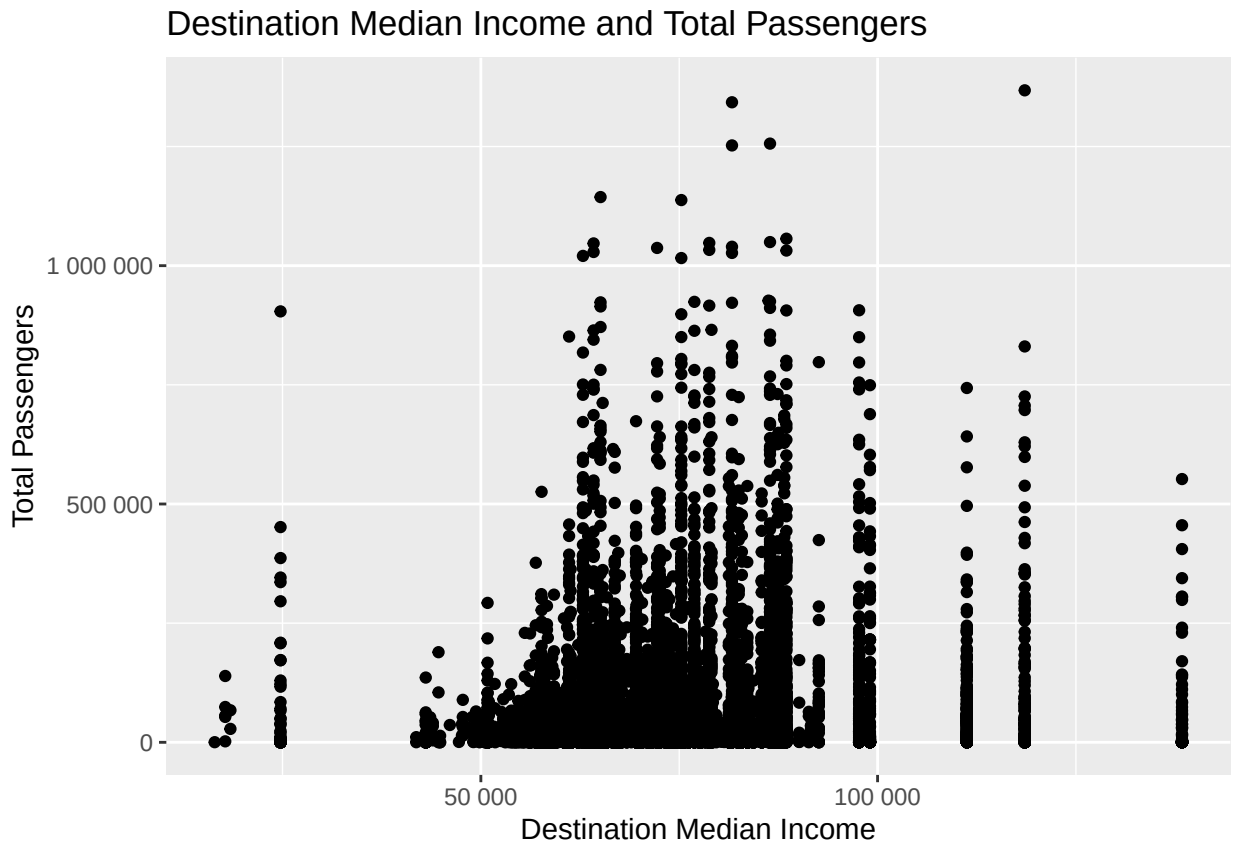

Destination Population and Total Passengers
Total Passengers
1 000 000
500 000
0
0
5 000 000
10 000 000
15 000 000
20 000 000
Destination Population




```
## Warning: Removed 98 rows containing missing values or values outside the scale range
## (`geom_point()`).
```




```
## Warning: Removed 99 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



When looking at Origin Population and Total Passengers, there isn't too much of a trend, but I do see a large obvious cluster in the bottom left, that suggests that flights coming out of lower-populated areas have fewer passengers, which makes sense. If an area is less populated, there are less people to buy those flights. We see roughly an identical trend for the Destination Population and Total Passengers graph as well. With Flight Distance and Total Passengers, I also see an obvious cluster in the lower left corner, with the majority of flights under 3,000. The flights with the most passengers are also on the left half of the graph. This suggests that very long flights have fewer passengers, and the flights with the most passengers are the shortest flights. This would make sense in reality- people make quick trips more often than they do very long flights, as it's rare for most people to travel internationally very frequently. When looking at Origin Median Income and Total passengers, we see a trend towards the middle of the map. This makes sense because the middle of the x-axis represents the most average income for people, which is where most common folk lie. There doesn't seem to be a strong trend when comparing the two axis though- if anything, we could say that the flights with the most passengers seem to trend towards the average income instead of outliers. The higher median income outliers still have many passengers, though. Lastly, when looking at Destination Median Income and Total passengers, we see a very similar trend with no major differences.

Question 3:

```
##
## Call:
## lm(formula = total_passengers ~ origin_population + dest_population +
##     distancemiles + origin_median_income + dest_median_income +
##     origin_per_capita_income + dest_per_capita_income + origin_median_home_value +
##     dest_median_home_value, data = cbsa_to_cbsa)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -313653  -58534  -30958    9413 1179058
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -6.922e+04  1.230e+04  -5.627 1.88e-08 ***
## origin_population           6.287e-03  3.291e-04  19.101  < 2e-16 ***
## dest_population             6.324e-03  3.306e-04  19.131  < 2e-16 ***
## distancemiles              -2.604e+01  2.023e+00 -12.869  < 2e-16 ***
## origin_median_income        1.606e+00  2.771e-01   5.794 7.11e-09 ***
## dest_median_income          1.534e+00  2.773e-01   5.532 3.25e-08 ***
## origin_per_capita_income -1.392e+00  5.137e-01  -2.710  0.00675 **
## dest_per_capita_income   -1.185e+00  5.135e-01  -2.309  0.02098 *
## origin_median_home_value -1.809e-02  1.269e-02  -1.425  0.15426
## dest_median_home_value   -1.906e-02  1.268e-02  -1.503  0.13293
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 127900 on 9250 degrees of freedom
##   (192 observations deleted due to missingness)
## Multiple R-squared:  0.108,  Adjusted R-squared:  0.1071
## F-statistic: 124.4 on 9 and 9250 DF,  p-value: < 2.2e-16
```

Interpreting Coefficients:

The coefficient for origin population is roughly 6.3, which means that for every 1,000 person increase in the population of the origin place, the total CBSA-to-CBSA passenger volume increases by around 6.3. The p-value is significant, less than 0.001, which makes sense in reality. Larger populations at an origin tend to generate more travel demand because there are naturally more people there to purchase these flights.

The coefficient for destination population is also around 6.3, which means that a 1,000 person increase in the population of the destination place is associated with an increase of around 6.3 passengers in travel volume. With a significant p-value of less than 0.001, this also makes sense because places with large populations attract more visitors and tourists.

When it comes to distance, the coefficient is around -2.6, meaning that for each additional mile of distance between the origin and destination, the passenger volume decreases by around 26 passengers. With a p-value of less than 0.001, this is also statistically significant. This also makes intuitive sense- longer flights usually have lower demand because of cost and travel time.

For origin media income, the coefficient is around 1.6, indicating that for each dollar increase in media income at the origin, there is an associated increase in passenger volume of about 1.6. With a statistically significant p-value of less than 0.001, this is intuitive- wealthier areas usually have more people who frequently travel as they are able to afford it.

With destination median income at 1.5, this means that each additional dollar in median income at the destination brings an increase of 1.5 passengers. This has a significant effect and indicates that destinations with a higher income could attract more visitors and tourists. For origin per capita income, the coefficient is -1.4, telling us that with a dollar increase in per capita income for the origin, the passenger volume increases by 1.4. Despite a significant p-value of 0.006, it doesn't seem intuitive.

On the flip side, for destination per capita income, the coefficient is -1.2, meaning that with a $1 increase in per capita income there is a decrease in about 1.2 passengers. Once again, despte a significant p-value, this also seems counterintuitive. When we look at median home value, the coefficient is -0.018, meaning that a dollar increase in the median home value at the origin means a decrease in about 0.01 passengers. This is not statistically significant with a p-value of 0.15.

Lastly, with median income in destinations, the coefficient s -0.019. A dollar increase in median household value at the destination brings a decrease of 0.019 passengers. This is also not statistically significant with a

p-value of 0.13.

Model Fit:

The R-squared value of 0.108 tells us that this model only explains about 10.8% of any differences in passenger volumes. While this value tells us that we have part of the story, it lets us know that there are several other factors that determine passenger volume that we did not take into consideration. Potential ideas for these variables may be seasonal patterns or holidays, weather, geographic preferences, and more.

The majority of the coefficients were statistically significant, which tells us that they likely do contribute to predicting passenger volume, namely population, income, and distance for both destination and origin. Median home values did not seem to impact the demand of passengers in this model.

Question 4:

```
##   origin_cbsa dest_cbsa distancemiles predicted_demand    route_direction
## 1         RDU       PDX          2363         31739.56 RDU to Destination
## 2         RDU       ELP          1606         20039.35 RDU to Destination
## 3         RDU       TLH           496         43697.81 RDU to Destination
## 4         RDU       SMF          2345         32657.93 RDU to Destination
## 5         PDX       RDU          2363         31696.98 Destination to RDU
## 6         ELP       RDU          1606         21409.93 Destination to RDU
## 7         TLH       RDU           496         43923.61 Destination to RDU
## 8         SMF       RDU          2345         33232.06 Destination to RDU
```

Based on the projected demand figures, the most popular routes are Raleigh-Durham to Tallahassee with a predicted demand of 43,698 passengers. Second, not far behind, is Tallahassee to Raleigh-Durham with a predicted demand of 43,924 passengers.

We would recommend these two routes for the new routes to be implemented. It is the quickest flight out of the four, which is in-line with our model indicating that shorter flights have higher demand. It is in the top ten cities in terms of population in Florida, which makes sense. Our model calculated that higher-populated areas have higher demand in passengers. There are higher-income areas in Tallahassee, but it is not necessarily a metropolitan hub, which could suggest some overestimation or error in our model.

Given the moderate to low R-squared value of 10.8%, our model does not explain the whole story behind passenger demand. This means that it would be a mistake to take the recommendation without considering other factors that could impact passenger demand. However, many of our coefficients were statistically significant, so it is safe to say that our model plays a role in predicting passenger volume, even if it is a smaller amount. We would recommend taking into account our recommendation but proceeding with caution as well as investigating other variables as well.