PLAN 372 Homework 4

Julia Bright

2025-03-31

R Markdown file and code can be found at: https://github.com/JBrighttt/plan372__hmks

```
library(tidyverse)
library(tidycensus)
rm(list=ls())
census_api_key("34511bfdd8cacadb68a5588d1847abdcc3defeea", install = TRUE,
↪  overwrite=TRUE)
airportpairs <- read_csv("plan372_hmks/hw4/airport_pairs.csv")
```

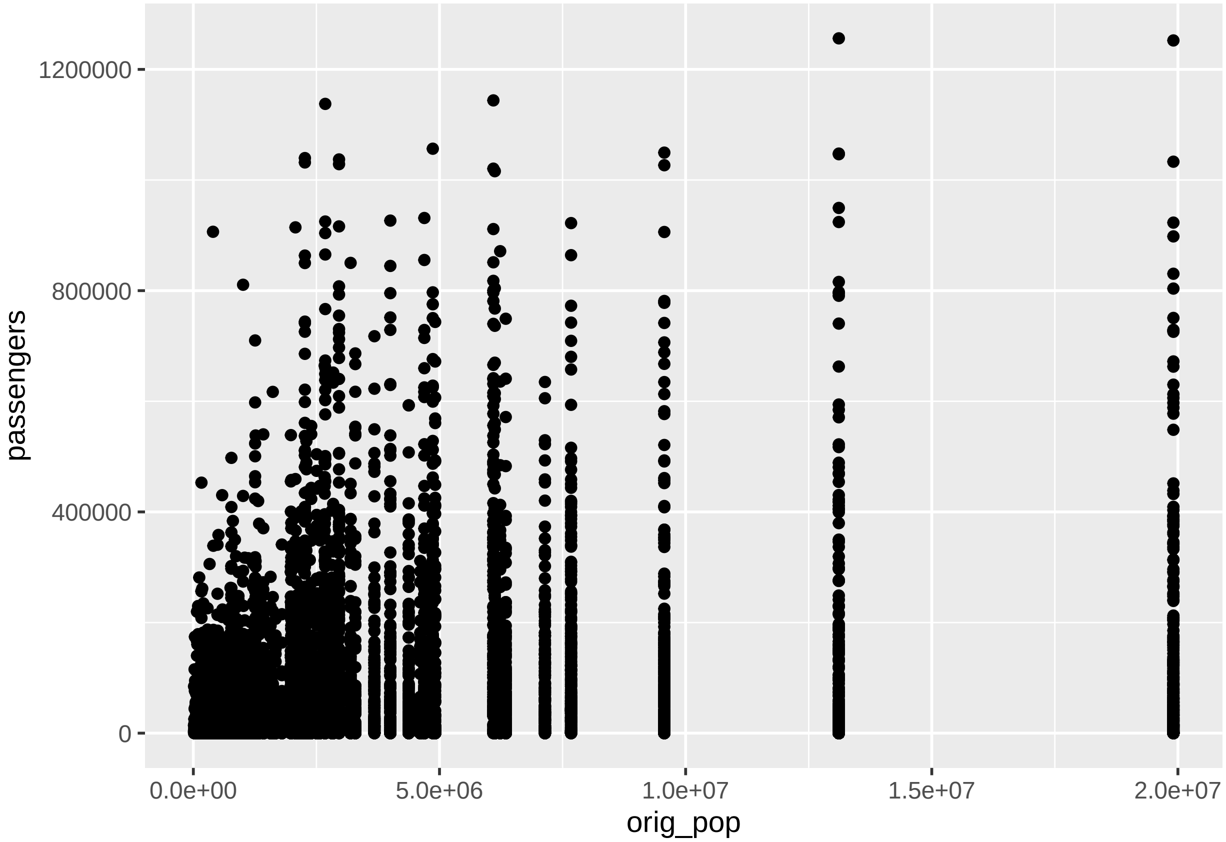
Question 1

```
marketsat <- airportpairs %>% # create dataframe with flight routes to and from RDU, with
↪  >= 10k passengers
  filter(origin=="RDU" | dest=="RDU") %>%
  filter(passengers>=10000)
```

Question 2

```
census_api_key("34511bfdd8cacadb68a5588d1847abdcc3defeea")
vars <- load_variables(2017, "acs5", cache = TRUE) #load variables to find the ones i
↪  need
acs_pop <- get_acs(geography="cbsa", year=2022, variables = c(population = "B01003_001"))
↪  #save dataframe with population from ACS
acs_income <- get_acs(geography = "cbsa", year=2022, variables = c(medianincome =
↪  "B19013_001")) #now with income

acs_income$GEOID <- as.numeric(acs_income$GEOID)
acs_pop$GEOID <- as.numeric(acs_pop$GEOID) #make numeric so it can join to cbsa values in
↪  airport pairs

#make copies of ACS data for joining population values
acs_orig <- acs_pop %>%
  rename(orig_pop = estimate) %>%
  rename(origin_cbsa = GEOID)

acs_dest <- acs_pop %>%
  rename(dest_pop = estimate) %>%
  rename(dest_cbsa = GEOID)

#make copies of income to join with orig and dest values
income_orig <- acs_income %>%
  rename(orig_income = estimate) %>%
  rename(origin_cbsa = GEOID)

income_dest <- acs_income %>%
  rename(dest_income = estimate) %>%
  rename(dest_cbsa = GEOID)

all <- full_join(airportpairs, acs_orig, by = "origin_cbsa")
all <- full_join(all, acs_dest, by = "dest_cbsa")
all <- full_join(all, income_orig, by = "origin_cbsa")
all <- full_join(all, income_dest, by = "dest_cbsa")
#join data, remove redundant and not-useful variables
all <- all %>%
  mutate(moe.x = NULL) %>%
  mutate(moe.y = NULL) %>%
  mutate(variable.x = NULL) %>%
  mutate(variable.y = NULL) %>%
  mutate(NAME.x = NULL) %>%
  mutate(NAME.y = NULL) %>%
  mutate(moe.x.x = NULL) %>%
  mutate(moe.y.y = NULL) %>%
  mutate(variable.x.x = NULL) %>%
  mutate(variable.y.y = NULL) %>%
  mutate(NAME.x.x = NULL) %>%
  mutate(NAME.y.y = NULL)

all_cbsa <- all %>% #create dataset of cbsa to cbsa to remove multiple airports serving
↪   one cbsa
  group_by(origin_cbsa_name, dest_cbsa_name) %>%
  summarize(passengers, distancemiles, orig_pop, dest_pop, orig_income, dest_income)
```

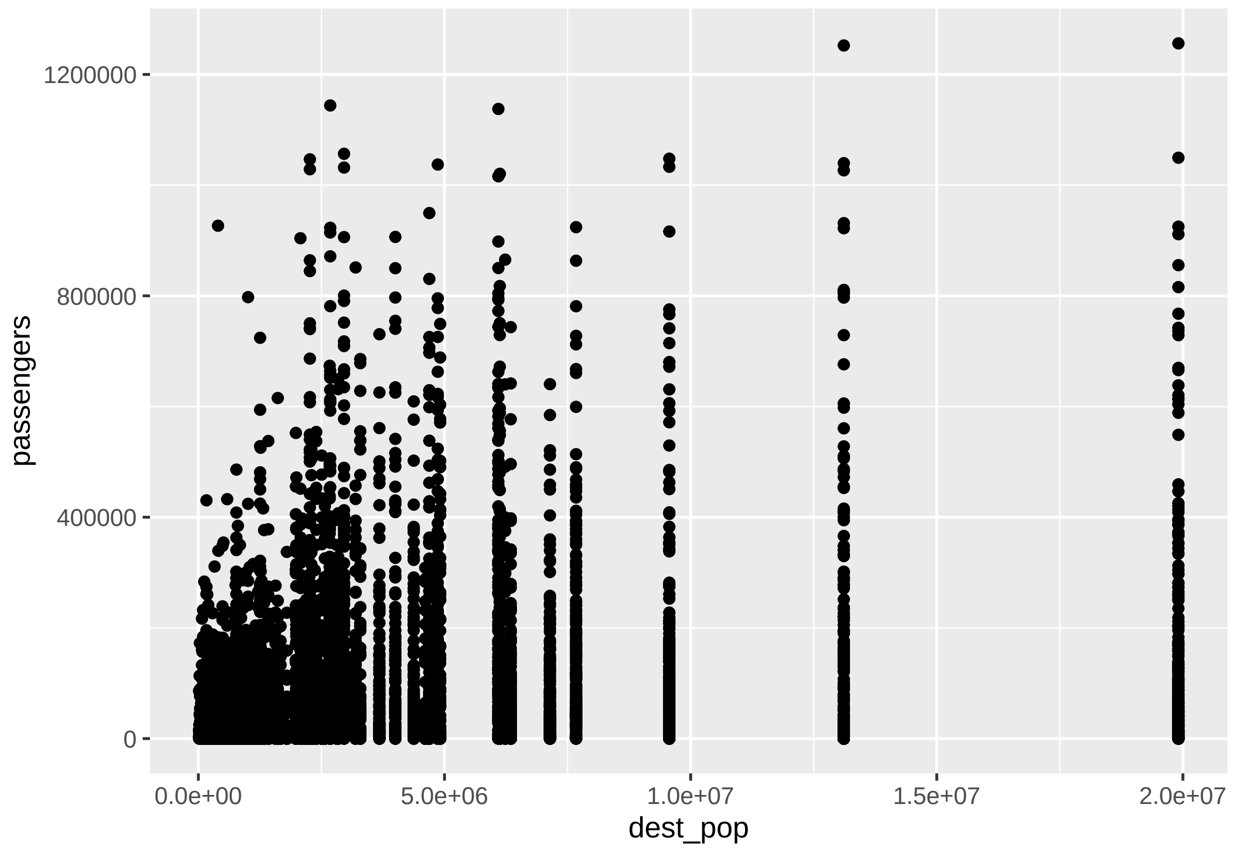
```
ggplot(all_cbsa, aes(x=orig_pop, y=passengers)) + geom_point() #scatterplot origin pop
↪   and passengers
```



```
#people are going, from each city of origin, that city's population measured on X axis
```

There appears to be a general trend that cities with higher populations trend higher in the amount of outbound flights.

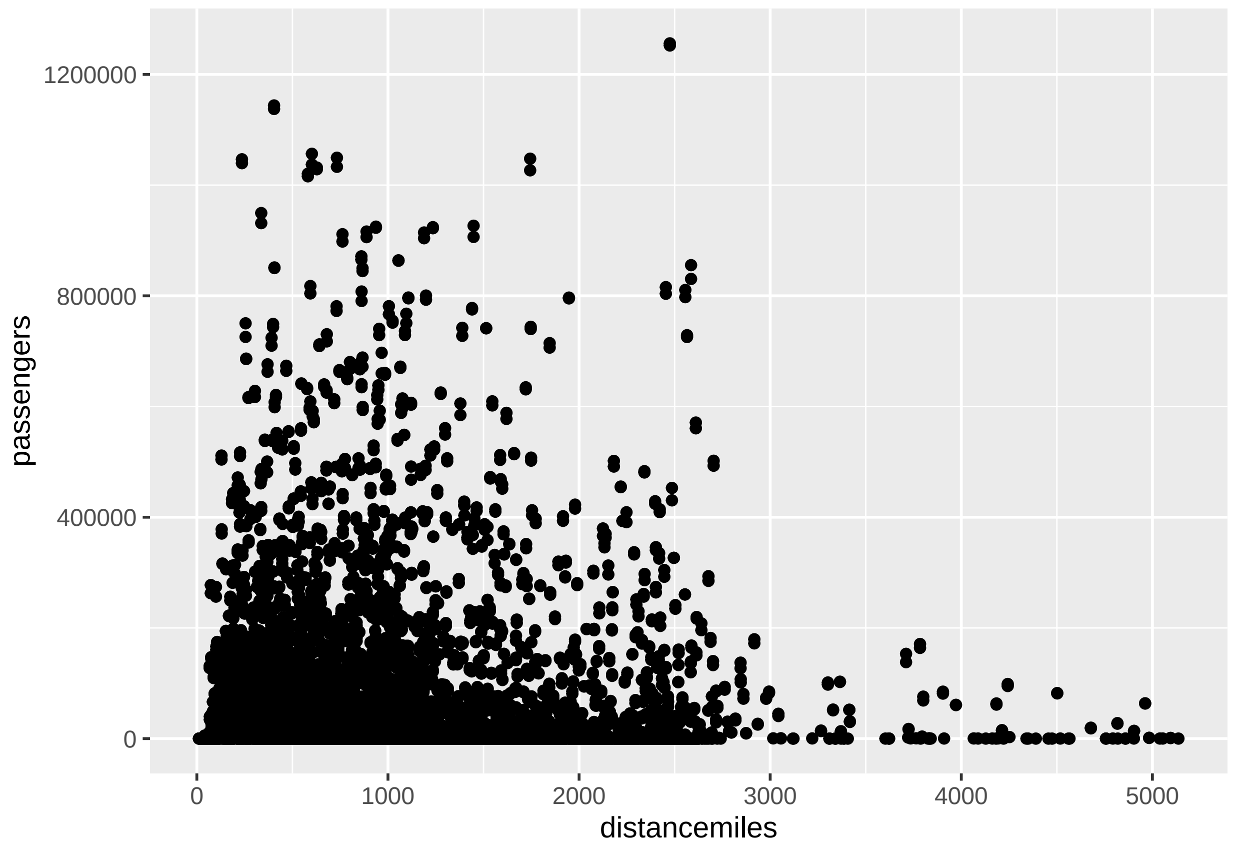
```
ggplot(all_cbsa, aes(x=dest_pop, y=passengers)) + geom_point() #scatterplot destination
↪  pop and passengers
```



```
#people are coming, to city of destination, that city's population measured on X axis
```

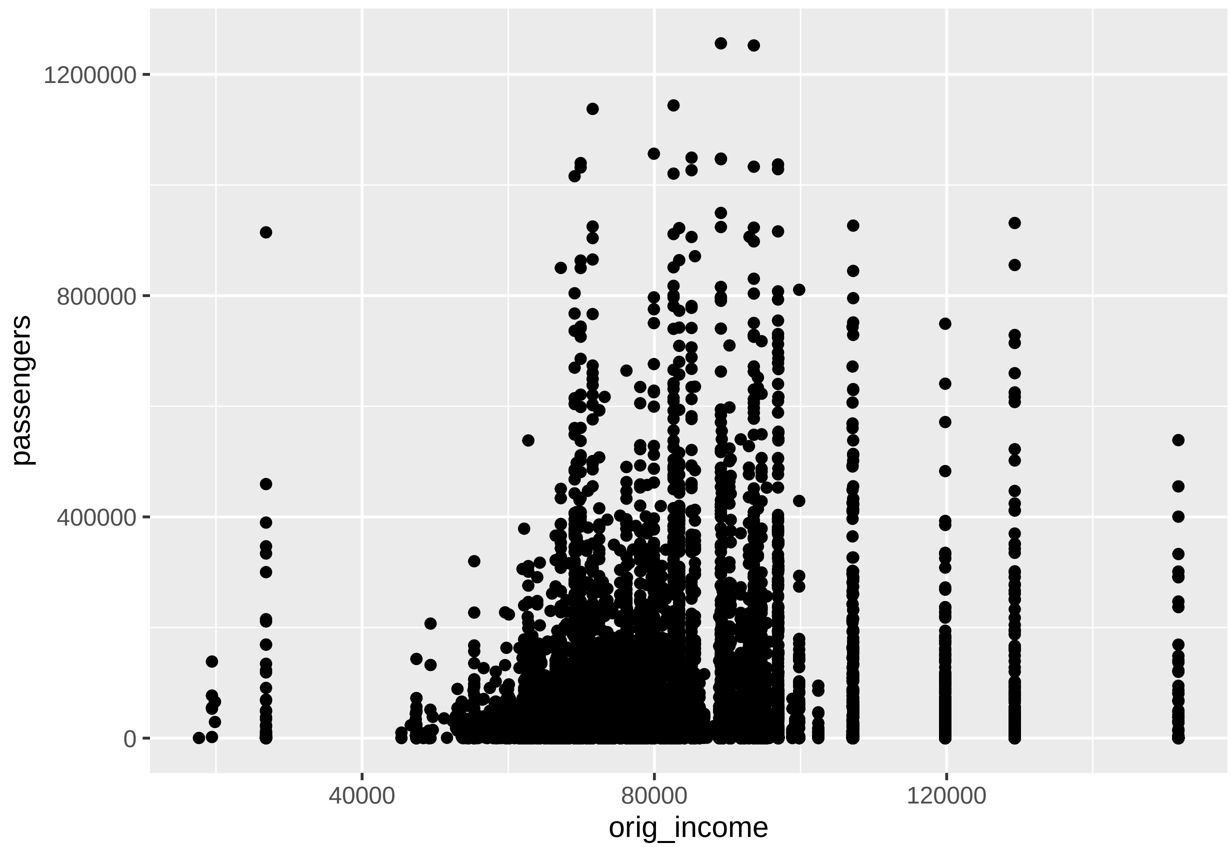
There appears to be a similar trend when looking at the population of the destination cities: larger cities have more inbound flights.

```
ggplot(all_cbsa, aes(x=distancemiles, y=passengers)) + geom_point() #scatterplot flight
↪  distance and passengers
```



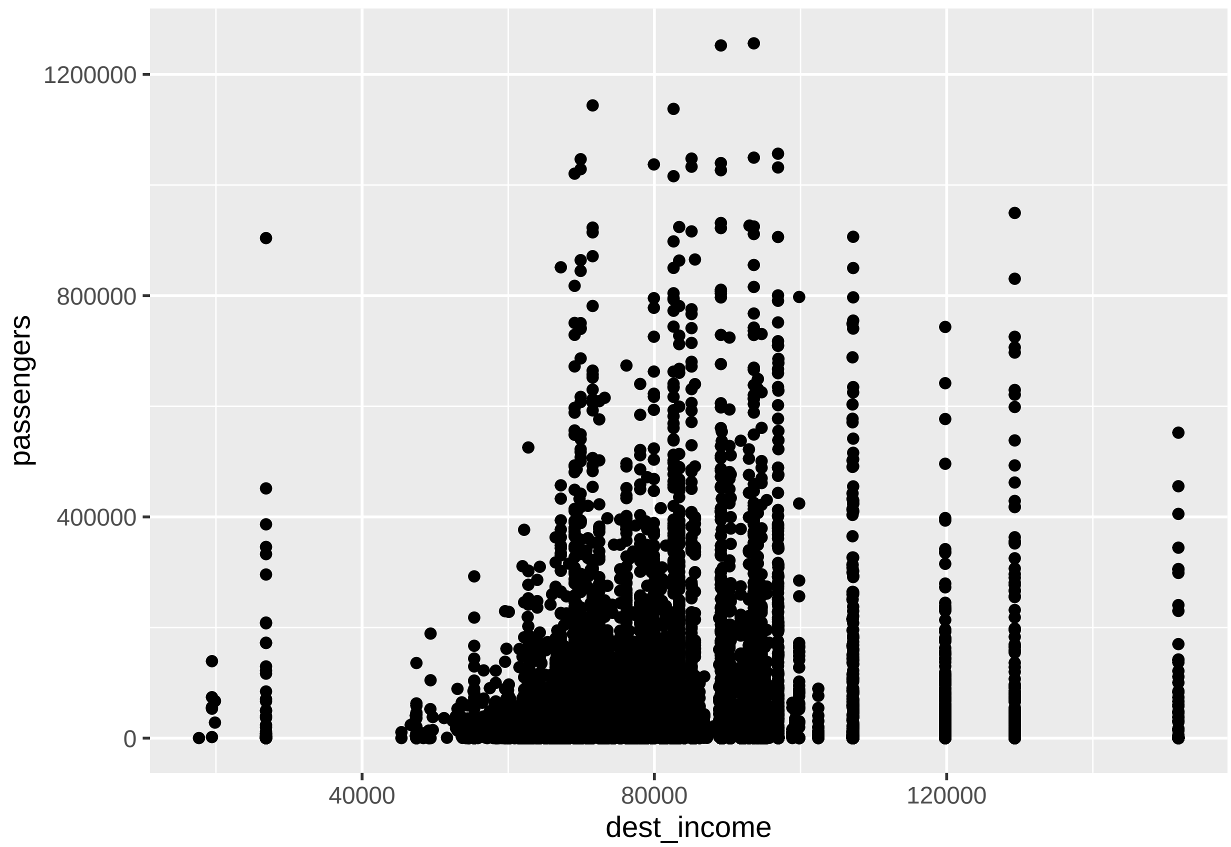
There appears to be a negative relationship between flight distance and number of passengers; the longer a flight, the lower the demand.

```
ggplot(all_cbsa, aes(x=orig_income, y=passengers)) +geom_point() #scatterplot origin
↪  income and passengers
```



It appears that the demand for outbound flights are highest in cities with incomes averaging around 60k to 100k. It seems to be a generally linear relationship, until reaching around 100k.

```
ggplot(all_cbsa, aes(x=dest_income, y=passengers)) +geom_point() #scatterplot destination
↪  income and passengers
```



A similar distribution shown here, with the majority of passengers heading to cities around the same income brackets.

Question 3

```
library(broom)
fit1 <- lm(passengers~ orig_pop + distancemiles + orig_income, data=all_cbsa) # linear
↪  model regressing passenger count on origin population and income
fit_orig <- tidy(fit1)
fit2 <- lm(passengers~ dest_pop + distancemiles + dest_income, data=all_cbsa) # linear
↪  model regressing passenger count on destination population and income
fit_dest <- tidy(fit2)

fit_orig
```

```
## # A tibble: 4 x 5
##   term             estimate   std.error statistic  p.value
##   <chr>               <dbl>       <dbl>     <dbl>    <dbl>
## 1 (Intercept)   22241.      6727.            3.31 9.48e- 4
## 2 orig_pop          0.00485    0.000321     15.1  5.57e-51
## 3 distancemiles   -19.2        1.90        -10.1  7.58e-24
## 4 orig_income       0.614      0.0841        7.30 3.07e-13
```

```
#Coefficients:
#(Intercept)       orig_pop  distancemiles    orig_income
#     22240        0.004849     -19.16           0.6144
#       Per 1000:  4.849
# All coef's statistically significant, p<.05
# Residual standard error: 130400 on 9395 degrees of freedom
#  (1345 observations deleted due to missingness)
# Multiple R-squared:  0.04626, Adjusted R-squared:  0.04596
```

Based on this linear model, a 1000-person increase in population at the origin city is associated with a 4.849 passenger increase in outbound flights. Every increase of 1 mile in flight distance is associated with a reduction of 19.16 passengers. Finally, each increase of 1 dollar in average income in a metropolitan area is associated with a .6144 increase in passengers. All coefficients are statistically significant. However, with a very low R-squared of 0.046, the model may not be a great fit for the data.

```
fit_dest
```

```
## # A tibble: 4 x 5
##   term             estimate   std.error statistic  p.value
##   <chr>               <dbl>       <dbl>     <dbl>    <dbl>
## 1 (Intercept)   21541.      6748.            3.19 1.42e- 3
## 2 dest_pop          0.00489    0.000322     15.2  1.96e-51
## 3 distancemiles   -19.3        1.90        -10.2  3.03e-24
## 4 dest_income       0.623      0.0845        7.37 1.82e-13
```

```
#Coefficients:
#  (Intercept)       dest_pop  distancemiles    dest_income
#    21540          0.004894     -19.33           .6227
#     Per 1000:     4.894
# All coef's statistically significant, p<.05
# Residual standard error: 130400 on 9394 degrees of freedom
#  (1346 observations deleted due to missingness)
#Multiple R-squared:  0.04675,  Adjusted R-squared:  0.04645
```

Based on this linear model, a 1000-person increase in population at the destination city is associated with a 4.894 passenger increase in outbound flights. Every increase of 1 mile in flight distance is associated with a reduction of 19.33 passengers. Finally, each increase of 1 dollar in average income in a metropolitan area is associated with a .6227 increase in passengers. These are all slightly greater effects when looking at the destination city of a flight, rather than the origin city. This could suggest that factors such as population or average income could create more passenger demand in a destination city compared to an origin city. All coefficients are statistically significant. However, with a very low R-squared of 0.046, the model may not be a great fit for the data.

Question 4

```
newdata <- data.frame(Destination = c("Portland", "El Paso", "Tallahassee",
↪  "Sacramento"), distancemiles = c(2363, 1606, 496, 2345), orig_pop = c(2505312,
↪  867161, 386064, 2394673), dest_pop = c(2505312, 867161, 386064, 2394673), orig_income
↪  = c(90451, 55344, 59757, 89227), dest_income = c(90451, 55344, 59757, 89227))  #
↪  create new dataframe for four airports being considered by Air Carolina
```

```
newdata$pv_inbound<- predict(fit1, newdata = newdata) # prediction of demand by origin
↪  population and income
newdata$pv_outbound<- predict(fit2, newdata = newdata) # prediction of demand by
↪  destination population and income

newdata
```

```
##   Destination distancemiles orig_pop dest_pop orig_income dest_income
## 1    Portland          2363  2505312  2505312       90451       90451
## 2     El Paso          1606   867161   867161       55344       55344
## 3 Tallahassee           496   386064   386064       59757       59757
## 4  Sacramento          2345  2394673  2394673       89227       89227
##   pv_inbound pv_outbound
## 1   44697.91    44448.56
## 2   29685.22    29203.49
## 3   51327.59    51053.88
## 4   43754.16    43492.87
```

Raleigh to Tallahassee appears likely to have the most demand, based on this predictive model. Once again, with a very low R-squared of 0.046, this model cannot be taken with great confidence. However, the higher demand relative to the other routes makes some sense, when considering that flight distance appears to have a considerable impact on passenger demand. The predicted demand of Raleigh to El Paso seems to stray from this expectation, being over 700 miles shorter than both flights to Portland and Sacramento, but having a much lower predicted demand.