You have been contracted by a group of investors interested in starting a new airline, Air Carolina, based at RDU airport. They feel that RDU does not have enough domestic non-stop flights, and want Air Carolina to fill that void. They want you to build a model to help them understand which routes are likely to have enough demand to support new non-stop service. They want to use the model to understand what factors are associated with a popular air route, and to forecast demand on some of the routes they are considering.

They have provided you with data on the popularity of existing US domestic air traffic by major carriers, derived from the United States Bureau of Transportation Statistics Airline Origin and Destination Survey DB1B database. This dataset includes total passenger numbers for October 2021 - September 2022 between all airport pairs in the US served by nonstop flights. The data also include information on what Census [metropolitan or micropolitan areaLinks to an external site.](https://www.census.gov/programs-surveys/metro-micro.html) each airport is a part of, and the distance between each pair of airports. You can use the tidycensus package to retrieve demographic information about each metropolitan area, to include as demographic variables in your model. You can choose the variables you think are best, but you should definitely include population. You may also want to look at economic indicators—for example, high-income population, as higher-income people fly more often.

This report should be uploaded to Canvas, and should include a link to your code on Github (repository can be public or private; if private, add me as a collaborator - my github username is arelkhattabi). Feel free to use Quarto, Markdown, Canva, or any other software to create your report.

You should describe your findings and recommendations in a well-written report to Air Carolina investors (complete sentences, please 🙂).

In addition to the points below, you will get:

* 1 point for code uploaded to Github, with comments explaining what each line does
* 1 point for code that runs top to bottom without errors (If you don’t get this point but the code runs on your machine without errors, talk to me in class or office hours. If the error is due to differences between our machines, I will give you this point back.)

Question 1: Market saturation analysis

The first question the investors want to understand is how popular the existing routes from or to RDU are. Create a table of the existing flights to or from RDU, and the number of passengers passenger traveling to each destination. Make sure to include both flights departing RDU and those arriving RDU. There are a few records in the data for flights between RDU and places that do not have nonstop service from RDU (e.g. Fairbanks, Tucson). Filter your table to only include airport pairs with more 10,000 passengers. [0.5 points]

Question 2: Bringing in Census data

Use tidycensus to retrieve Census data and match it to the airport pair data (the full dataset, not only flights to/from RDU). The geography you’ll want to use with tidycensus is cbsa for core-based statistical area; this will match the origin\_cbsa and dest\_cbsa codes in the airport data. Core-based statistical area is Census jargon for a city and its suburbs. Retrieve at least the total population, but consider whether there are any other variables you want to include that may be predictive of air travel volumes.

Join the CBSA population data to the airport data. You will need to join it twice, so that you have total population for the origin CBSA and the destination CBSA. You may want to create copies of the Census data with the variables renamed before joining, so that you know which columns are for the origin vs. destination airport. For instance, you might create one table with the population column called “origin\_pop” and then join based on the origin CBSA, and another with the column called “destination\_pop”, and then join based on the destination CBSA. If you get errors about types not matching, look at how we solved this in the R GIS exercise.

Some areas are served by multiple airports (for instance, Chicago is served by both O’Hare and Midway International Airports). The data show airport-to-airport volumes. Use group\_by and summarize to create a new dataset that shows total CBSA to CBSA volumes.

A few airports are not in a metropolitan area. You can remove them from this analysis.

Generate scatterplots between origin population and total passengers, destination population and total passengers, and flight distance and total passengers. Include the scatterplots in your report, and describe what trend they show (if any). [1.5 points]

*Extra credit*: include a pair of scatterplots for another variable other than population, at the origin and destination [+1 point]

Question 3: Passenger volume regression

Run a regression of total CBSA-to-CBSA passenger volumes on origin CBSA population, destination CBSA population, distance between cities, and any other Census variables you wish to include. Use all the data, not just data for flights coming to/going from RDU. Present the results of your regression as a table in your writeup, and interpret the coefficients (e.g., how many more airline passengers are associated with a 1000-person increase in population at the origin city, etc.) Interpret all coefficients in the model (except the constant/intercept). Do the coefficients make sense? Are they statistically significant? Does the model fit well, as indicated by R2? [2 points + 1 extra credit point for including and interpreting variables other than population]

Question 4: Passenger volume prediction

These are the routes Air Carolina is considering implementing:

* RDU to Portland, OR (PDX) - 2,363 mi
* RDU to El Paso, TX (ELP) - 1,606 mi
* RDU to Tallahassee, FL (TLH) - 496 mi
* RDU to Sacramento, CA (SMF) - 2,345 mi

Use your model to predict passenger demand on these routes. Be sure to predict demand in both directions, i.e. RDU to the destination, and the destination to RDU. Since your model is based on origin and destination metropolitan area Census data, you’ll have to find that data for these cities, and create a new table to predict from (it should already be in your dataset, since airlines already fly to all of these cities, just not from RDU).

Include a table with forecast demand for each route in your writeup, and discuss which route is likely to be the most popular. How confident are you in your projection, based on the R2 of your model? Are there any unexpected or nonsensical results? [2 points]

Additional background information (optional, for your curiosity only)

The model you built above is a simple form of a [spatial interaction modelLinks to an external site.](https://researchrepository.wvu.edu/cgi/viewcontent.cgi?article=1010&context=rri-web-book). More complex spatial interaction models use a *gravity* formulation, which models interaction between two places using the same or similar formulae as the formula used to determine the force of gravity between two objects:

f=s1s2d2

where f is the force or level of spatial interaction between two objects, s1 and s2 are the sizes of the two objects (mass for gravity, might be population for example in a general spatial interaction model). d is the distance between the two objects. More general spatial interaction models allow the exponent 2 to vary, and may also add exponents to the size terms:

i=s1λs2αdβ

(see page 12 of [Fotheringham and HanesLinks to an external site.](https://researchrepository.wvu.edu/cgi/viewcontent.cgi?article=1010&context=rri-web-book))

We didn’t talk about variable transformations in regression, but one very useful transformation is the logarithm, which converts multiplication into addition. By taking the logarithm of both sides and applying [logarithmic identities](https://uncch.instructure.com/courses/68107/assignments/List_of_logarithmic_identities), we can re-write the equation like this:

lni=λlns1+αlns2−βlnd

which looks a lot like the equations we had above, and we can then use linear regression to estimate this model.

Spatial interaction models are used a lot. Most regions have a travel demand model which they use to forecast future demands on the transportation system. They use spatial interaction models like this to estimate where people will go in their day-to-day travel—people prefer places closer to them, but also will travel further to get to larger places.