Final Project

Ryan Yu

Introduction and data

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.3     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.0
v ggplot2   3.4.3     v tibble    3.2.1
v lubridate 1.9.2     v tidyr     1.3.0
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
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 becom
-- Attaching packages --------------------------------------- tidymodels 1.1.1 --

v broom        1.0.5     v rsample      1.2.0
v dials        1.2.0     v tune         1.1.2
v infer        1.0.5     v workflows    1.1.3
v modeldata    1.2.0     v workflowsets 1.0.1
v parsnip      1.1.1     v yardstick    1.2.0
v recipes      1.0.8

-- Conflicts ------------------------------------------ tidymodels_conflicts() --
x scales::discard() masks purrr::discard()
x dplyr::filter()   masks stats::filter()
x recipes::fixed()  masks stringr::fixed()
x dplyr::lag()      masks stats::lag()
x yardstick::spec() masks readr::spec()
x recipes::step()   masks stats::step()
* Use tidymodels_prefer() to resolve common conflicts.

New names:
Rows: 25976 Columns: 25
```

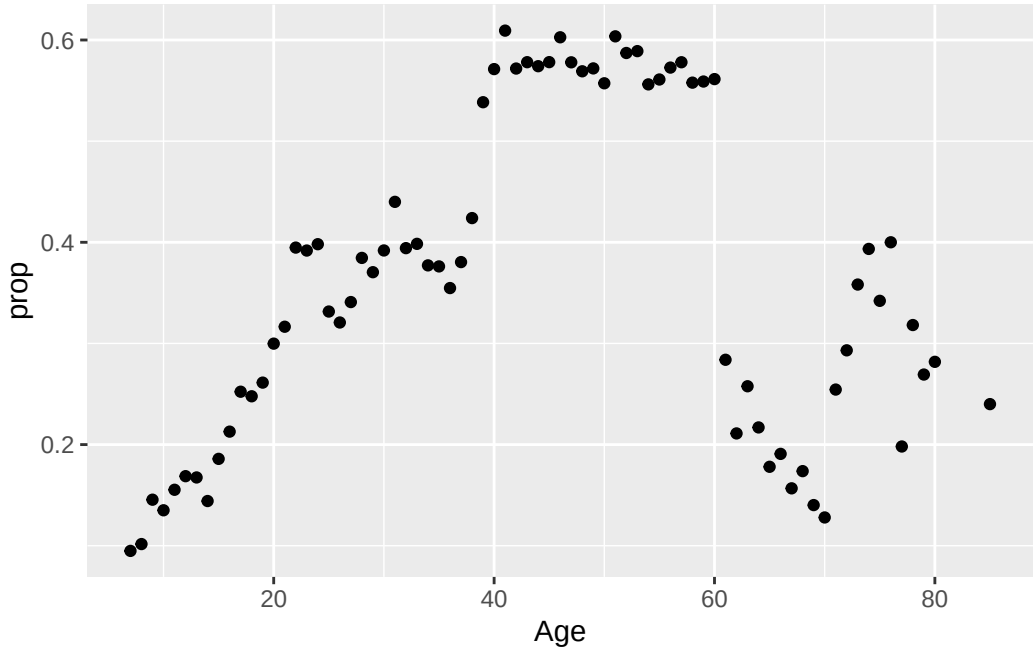
```
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (5): Gender, Customer Type, Type of Travel, Class, satisfaction
dbl (20): ...1, id, Age, Flight Distance, Inflight wifi service, Departure/A...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
New names:
Rows: 103904 Columns: 25
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (5): Gender, Customer Type, Type of Travel, Class, satisfaction
dbl (20): ...1, id, Age, Flight Distance, Inflight wifi service, Departure/A...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
Joining with `by = join_by(...1, id, Gender, `Customer Type`, Age, `Type of Travel`, Class,
```

Air travel is an extremely popular form of transportation in the United States with over a million people flying every day. Everyone's experience is unique, whether they are travelling for work, vacation, school, etc. My research is motivated by the many different factors of air travel that contribute to a passenger's satisfaction. Understanding the factors that contribute to passenger satisfaction is crucial for airlines to improve their services. My research aims to answer the question: What are the most important factors that drive overall passenger satisfaction, and can we use these factors to predict customer satisfaction? My data set was collected through a US passenger satisfaction survey and compiled on Kaggle. The data set was originally split into "training" and "testing" data, which were random mutually exclusive parts of the same survey. In my research, the two data sets are recombined and an additional binary variable for satisfaction was added. Fourteen factors were included in the survey, with participants rating their satisfaction for each factor from 1 to 5, with 1 being least satisfied and 5 being most satisfied. Additionally, a rating of 0 corresponded to "Not Applicable", however I have changed the satisfaction levels of 0 to NA as to not skew the analysis. Other variables include gender, customer type (loyal/disloyal), age, type of travel (business/person), class of travel, flight distance, departure delay, arrival delay, and overall satisfaction.

```
# A tibble: 1 x 14
    iws  datc   eob    gl   fad    ob    sc   ife   obs   lrs    bh   cis   ifs
  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1  2.81  3.22  2.88  2.98  3.21  3.33  3.44  3.36  3.38  3.37  3.63  3.31  3.64
# i 1 more variable: c <dbl>
```



(will put this into a nicer looking graph later) The average satisfaction score varies between 2.81 for in-flight WiFi service to 3.64 for in-flight service. The probability of being satisfied seems highest for middle age (40-60)people with low satisfaction levels for younger and older people.

This section includes an introduction to the project motivation, data, and research question. Describe the data and definitions of key variables. It should also include some exploratory data analysis. All of the EDA won't fit in the paper, so focus on the EDA for the response variable and a few other interesting variables and relationships.

The research question and motivation are clearly stated in the introduction, including citations for the data source and any external research. The data are clearly described, including a description about how the data were originally collected and a concise definition of the variables relevant to understanding the report. The data cleaning process is clearly described, including any decisions made in the process (e.g., creating new variables, removing observations, etc.) The explanatory data analysis helps the reader better understand the observations in the data along with interesting and relevant relationships between the variables. It incorporates appropriate visualizations and summary statistics.