HW2

Lance Ding

2/21/2023

Instructions

This HW is worth 10 total points.

1. Make sure to load any packages you may need right at the start. Do *NOT* include the `learnr` package, ever, unless you are writing an interactive Tutorial (which you won't do in this course) - this will cause problems.

2. Ensure that no chunks have the `include = FALSE` or `echo = FALSE` option, as I want to be able to see *all* your code and output.

3. Brief but descriptive headings and document organization (answers under headings, text near relevant code, brief explanatory text as necessary, etc.) (1 pt)

   Look to my HW1 and RMarkdown Organization examples for how to write good headings, organize your assignment, and how much narrative text (outside of code chunks) I want. But a good rule of thumb is: **explain what YOU'RE doing, NOT what YOUR CODE is doing**. I do NOT need to see a repeat, line-by-line narration of what your code does - you can use code comments for that. (In complex blocks you usually want at least a comment every few lines.) I DO want to see an overall summary of what you did in your analysis.

4. Use the `flights` dataset from the `nycflights13` package again. While I didn't mention this in the Tutorial, these are only DOMESTIC flights - that is, those from NYC to another location within the U.S.

   i) Visualize the distribution of minutes flights spent in the air (`air_time`) a) overall and b) colored (or filled!) by `origin` (departing airport code - EWR for Newark, LGA for Laguardia, or JFK for John F. Kennedy) using your geom of choice (histogram, density, freqpoly, or ridgeplot/joyplot) in one or more code chunks. For the plot by airport, make sure we can see the distribution separately for each origin airport (that is, *don't* produce a stacked histogram).

      Give both plots a title, human-readable axis labels, and (if relevant) remove the legend title.

      Explain briefly what you see in narrative text. a) What is the shape of the distribution? Can you guess why it's shaped how it is (this requires some knowledge of U.S. geography, so especially if you're an international student it's OK if your guess is wrong)? b) Does the distribution of flight lengths vary any by airport? If so, what might explain that (HINT: not all airports have flights to all destinations)? (2 pts)

   ii) Visualize the number of flights per carrier and origin airport using a (stacked) bar chart. Color portions of the bar by airport.

      Give the plot a title, human-readable axis labels, and remove the legend title. **Color (fill) the bars using a viridis scale.**

      Explain briefly what you see in narrative text **by answering the following questions**. a) Which four airlines had the most total flights, and roughly how many for each? b) Did the majority of their flights depart from the same or different airports? Feel free to just identify the airline by their 2-letter code, but if you're curious which code corresponds to which airline you can look it up here. (2 pts)

5. Now use the `midwest` data frame included in the `tidyverse` package that provides some county-level demographic data for 5 Midwest states in the U.S.

   i) Provide a scatter plot illustrating the relationship between `percollege` (i.e., the percent of people who went to college in a county) and `percbelowpoverty` (i.e., the percent of that county living below the poverty line) with a smoothed fit line using the default smoothing method, LOESS.

      Summarize what you see in 1-2 sentences of narrative text. (1.5 pts)

   ii) Let's make this plot much prettier and more informative.

      a) Color the dots by `inmetro` (i.e., whether the county is in an urban metropolitan area (1) or a rural area (0)). Note you should use `color = as.factor(inmetro)` to make R treat `inmetro` as categorical rather than continuous.
      b) Facet the plot by state. Produce graphs with both fixed and free scales; decide which one you like best, and include that one in your report.
      c) Remove the smoothing lines to reduce clutter.
      d) Add an informative title, move the legend to the bottom, remove the legend title, and make the axis titles and legend labels informative and human-readable. Choose some non-default colors for the dots - use either the viridis colors or a palette appropriate for unordered categorical data from `RColorBrewer`. In short, pretty up the graph a bit.

      Summarize what you see in 3-4 sentences of narrative text **by answering the following questions**. a) Do urban or rural counties tend to have more or fewer people who went to college, and more or less poverty? b) Does the relationship between college and poverty differ in urban vs. rural counties (that is, does one rise or fall with the other in both types of counties)? c) Are these relationships consistent across states or not? d) Do you notice any interesting outlier counties? Pick at least one, go into the data, and name the county and state. Ideally from here you would do a little research to figure out *why* that county might be an outlier, but you'll still get full credit if you don't. (3.5 total pts)

Note: if you've done this properly, you should have *5* plots in the full report.

General organization and clarity of the report (e.g. headings, code in neat chunks, text written in appropriate spot) is worth 1 pt.

To submit this assignment:

Ideally, knit straight to PDF by changing `html_document` to `pdf_document` in line 5 above. Otherwise:

1. Knit to HTML. An HTML document should open automatically in another RStudio window.

2. Click "Open in Browser" in that HTML document. It should open as a webpage in your default browser (e.g. Chrome).

3. Click Ctrl+P/Command+P, but instead of printing a hard copy on your printer click "Save as PDF."

4. Save and upload that document to Canvas.

A note on PDF formatting: you may notice that long lines of code "fly off the side of the page" when you knit to PDF. To fix this:

If you're on a Windows machine:

- Install the `formatR` package

- Change your `opts_chunk$set` code line to the following: `knitr::opts_chunk$set(echo = TRUE, tidy.opts=list(width.cutoff=80), tidy=TRUE)`

That should force your code to always wrap rather than fly off the edge of the page of a PDF. Note this does not fix issues of, say, plot titles that are too long getting cut off. But it should fix all the errors with your code not wrapping. Happy PDFing!

If you're on a Mac: I don't have an easy solution for you. Try and keep your lines of code under about 80 characters. Feel free to use more vertical lines of code to accomplish this. But don't waste large amounts of time formatting. I'll ask you for clarification if something critical is missing.

——BEGIN ANSWER BELOW———

Lance Ding Homework 2: Analyzing `nycflights13` and `midwest` with Visualizations

Load the necessary packages

```
pacman::p_load(nycflights13, ggplot2, tidyverse, ggridges)
```

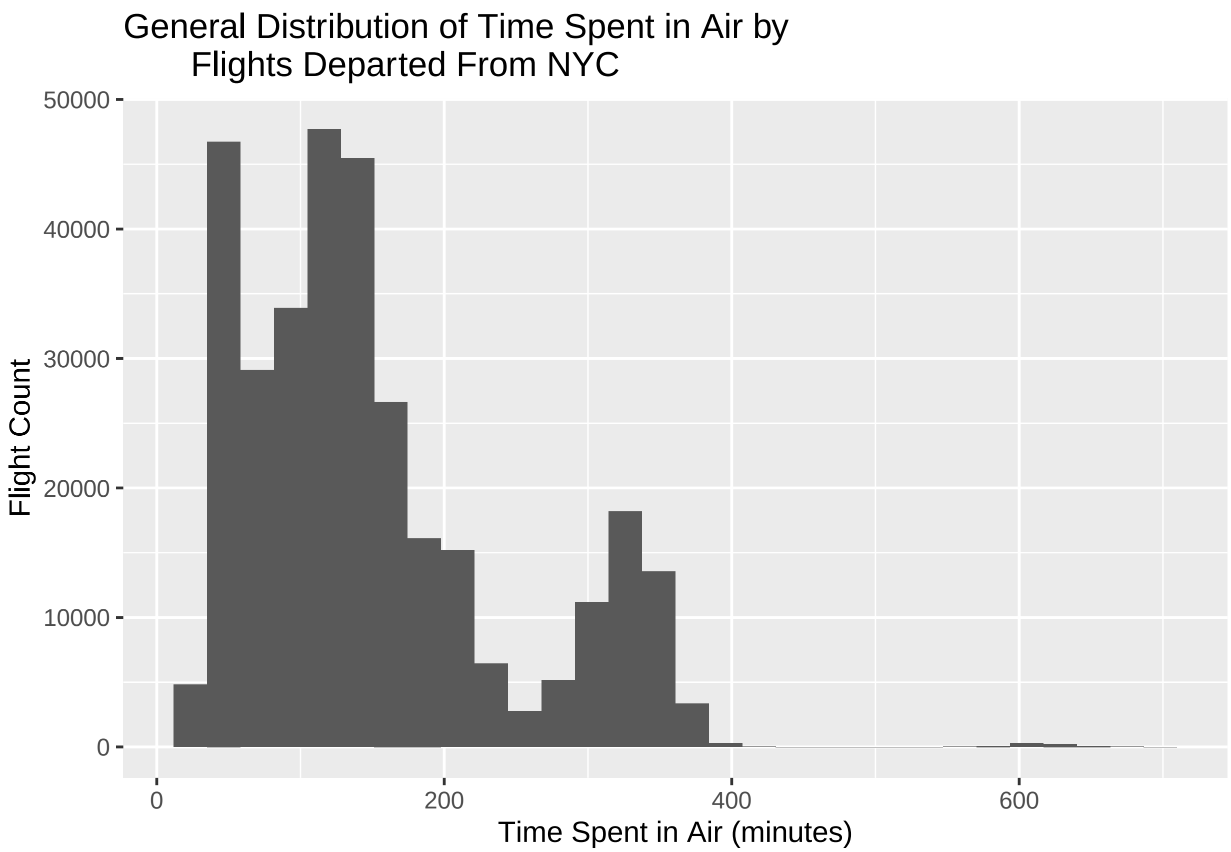
Q4i - Analyzing `air_time` Distribution

We are going to create a `ggplot` visualization of minutes of flight spent in the air (`air_time`) both **overall** and by **`origin`**. We will utilize histograms to showcase the distribution of `air_time` and use faceting for the plots of the respective airports of departure.

```
ggplot(data = flights, mapping = aes(x=air_time)) +
  geom_histogram() +
  labs(title = "General Distribution of Time Spent in Air by
       Flights Departed From NYC") +
  xlab("Time Spent in Air (minutes)") +
  ylab("Flight Count")
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```

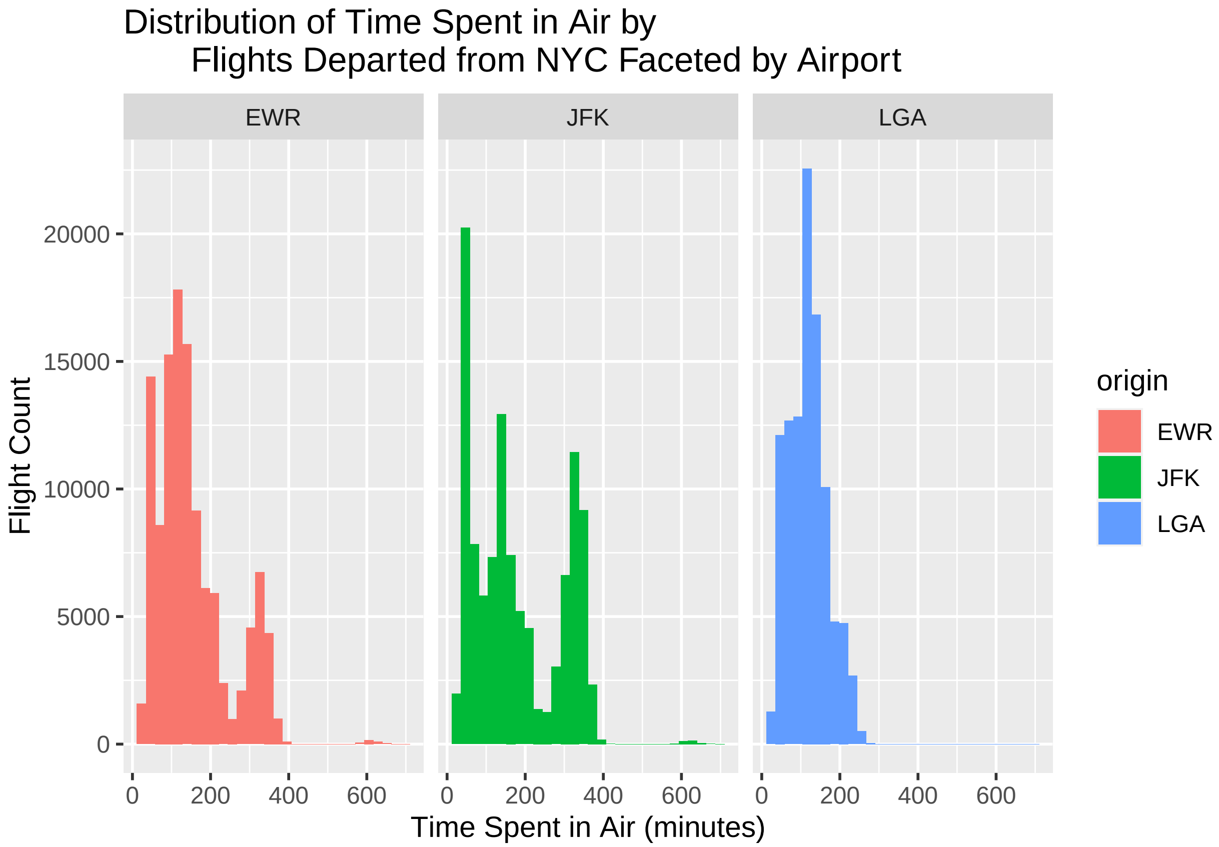
```
## Warning: Removed 9430 rows containing non-finite values (‘stat_bin()‘).
```



```
ggplot(data = flights, mapping = aes(x=air_time, fill=origin)) +
  geom_histogram() +
  facet_wrap(~origin) +
  labs(title = "Distribution of Time Spent in Air by
       Flights Departed from NYC Faceted by Airport") +
  xlab("Time Spent in Air (minutes)") +
  ylab("Flight Count")
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```

```
## Warning: Removed 9430 rows containing non-finite values (‘stat_bin()‘).
```

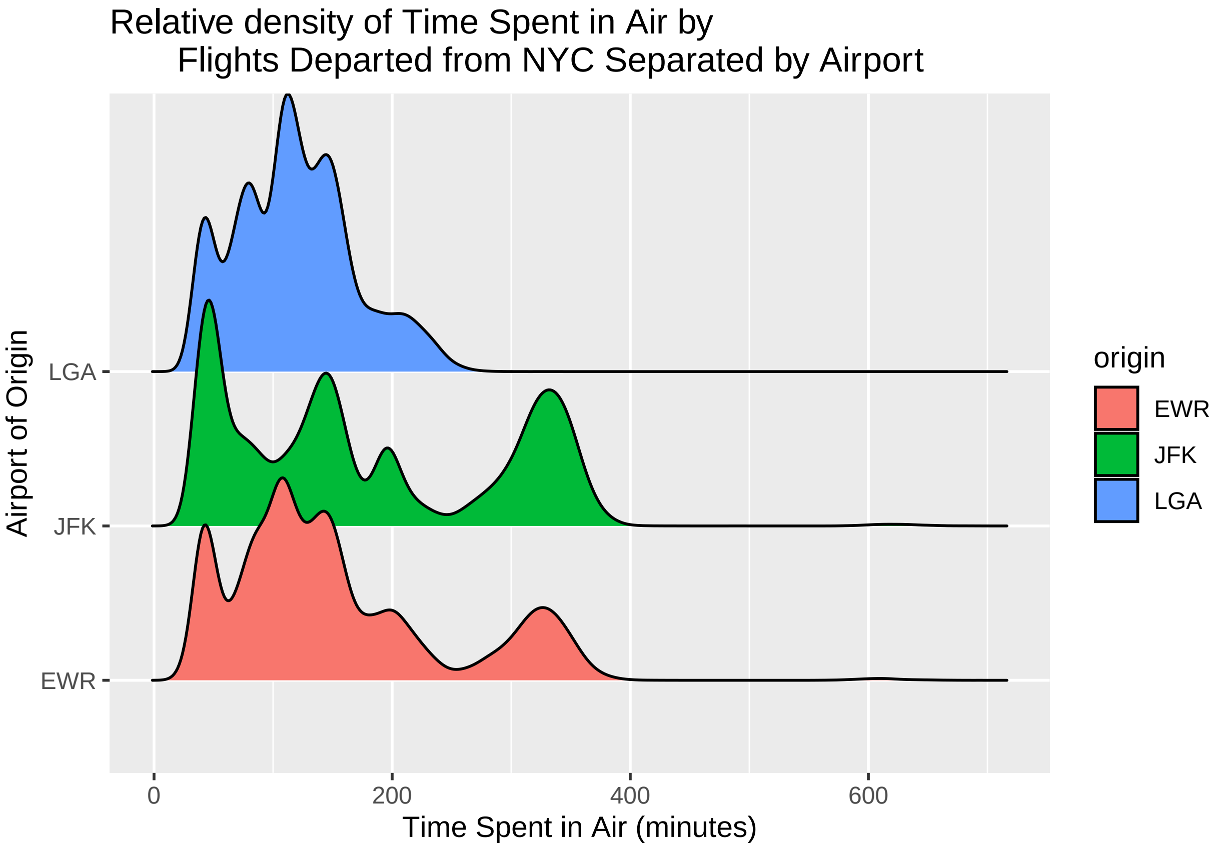


We might want another angle of the data with a vertically stacked view for other insights.

```
ggplot(data = flights, mapping = aes(x=air_time, y=origin, fill=origin)) +
  geom_density_ridges() +
  labs(title = "Relative density of Time Spent in Air by
       Flights Departed from NYC Separated by Airport") +
  xlab("Time Spent in Air (minutes)") +
  ylab("Airport of Origin")
```

```
## Picking joint bandwidth of 7.13
```

```
## Warning: Removed 9430 rows containing non-finite values
## (‘stat_density_ridges()‘).
```



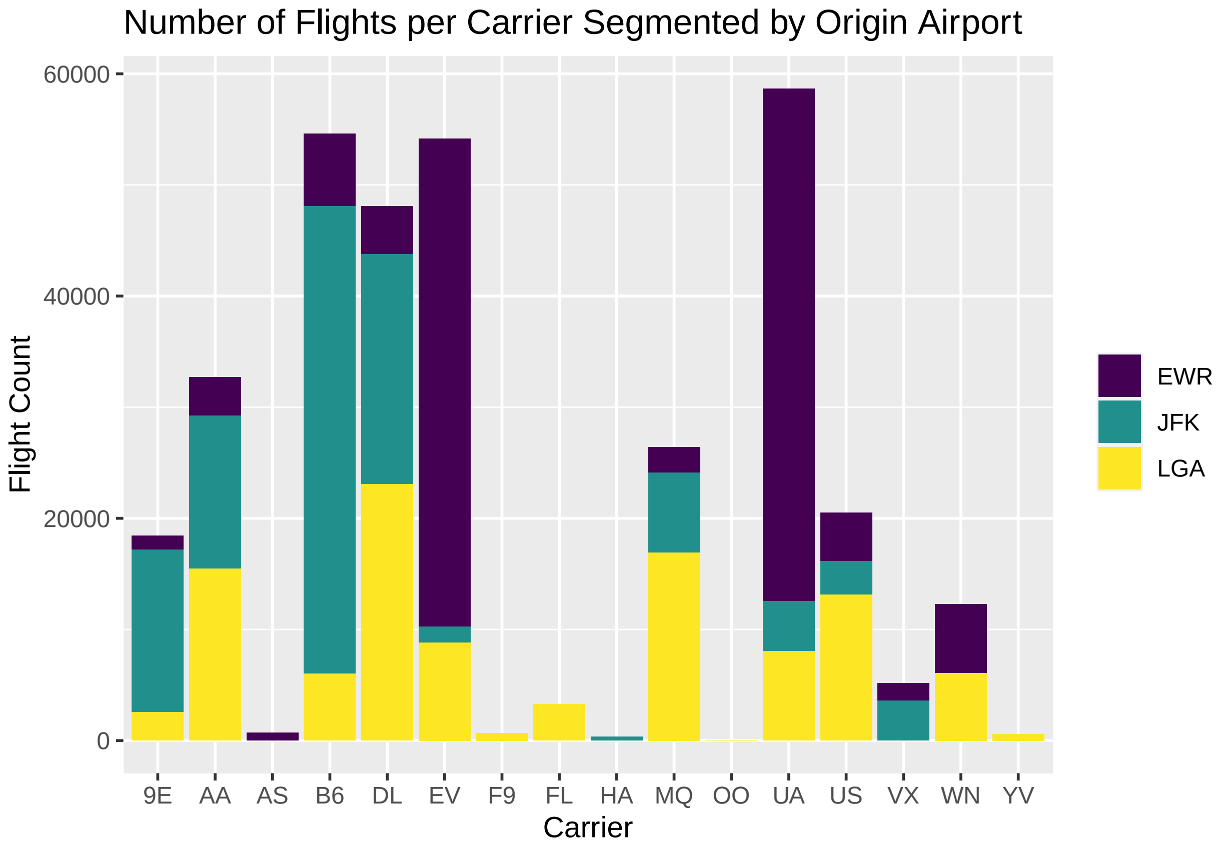
For the aggregate data, we have a somewhat bimodal distribution with a big peak around 150 minutes of flight time and a smaller peak at around 350 minutes of flight time. That translates to roughly 2.5 hours of flight time and 6 hours of flight time respectively. The major peak at 2.5 hours of flight time could potentially be explained by a high frequency of flights to *Chicago* and *Atlanta* (yay!), since flights from NYC to those regions take around 2.5 hours on average. The secondary peak at 6 hours of flight time could be explained by the flights going to the *west coast* - it takes around 6 hours to get to California and the west coast of Canada.

The distribution of flight lengths does vary by airport. Looking at the ridgeplot, we see that LGA (Laguardia) has most of their flights in a peak at lower ranges, while both JFK (John F. Kennedy) and EWR (Newark) have two distinct "regions" - a lower region for 2.5 hour and shorter flights and a region of 6 hour flights. LGA's lack of a second peak could be explained by their lack of flights to the west coast.

Q4ii - Visualizing Flights per Carrier and Origin Airport

We will now visualize the number of flights per carrier `carrier` and origin airport `origin` using a stacked bar chart, with the bar segmented by origin airport. Additionally, we will be utilizing the `scale_fill_viridis_d()` function to apply a viridis color scale to the bar segments. We choose to use the discrete option because we are dealing with a categorical variable in `origin`.

```
ggplot(data = flights, mapping = aes(x = carrier, fill = origin)) +
  geom_bar() +
  scale_fill_viridis_d() +
  labs(title = "Number of Flights per Carrier Segmented by Origin Airport") +
  ylab("Flight Count") +
  xlab("Carrier") +
  theme(legend.title = element_blank())
```



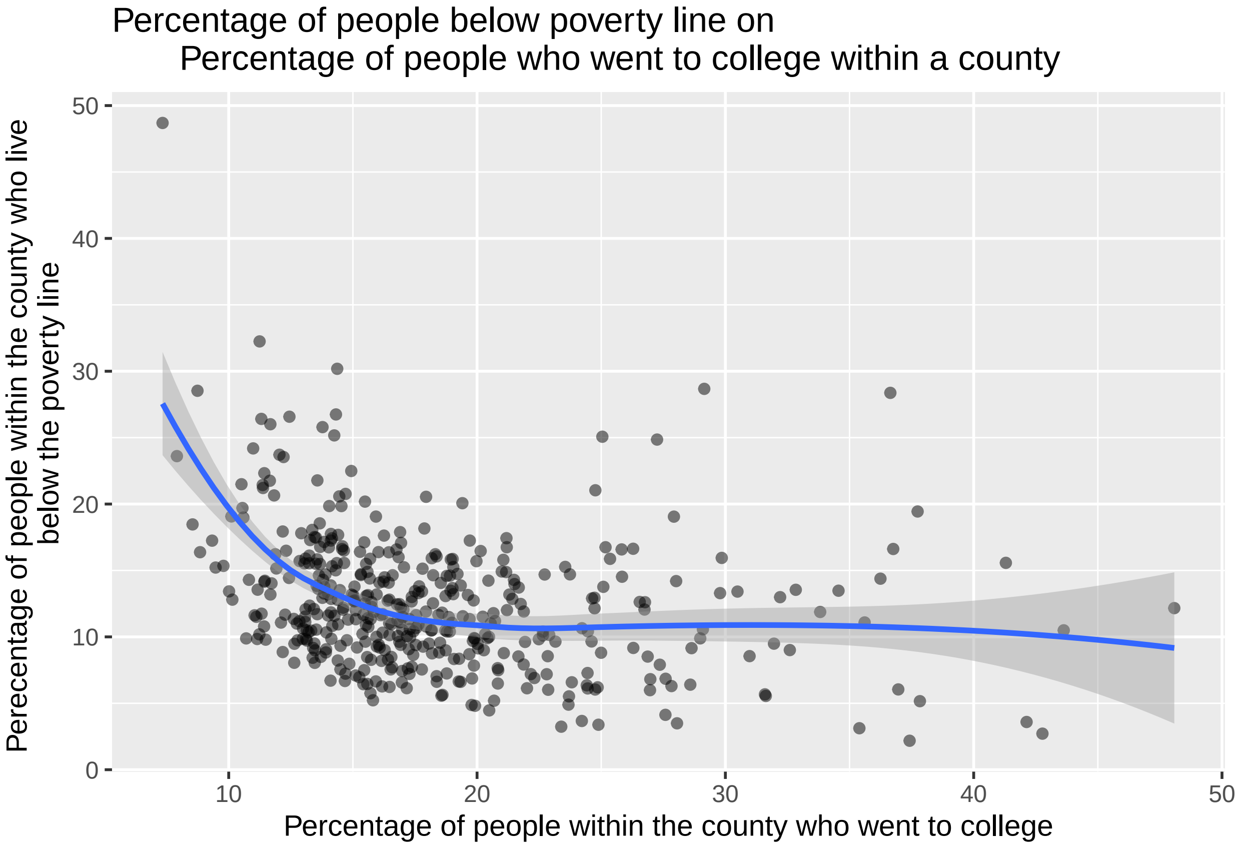
From these stacked bar charts, we can identify the **4** carriers with the most flights: **UA** with roughly 58000 flights, **B6** and **EV** both with roughly 55000 flights and **DL** with roughly 48000 flights. Looking at their respective bars, we can see that **UA** and **EV** have the majority of their flights coming out of EWR, **B6** with a majority of their flights coming out of JFK and **DL** with a majority of their flights coming out of JFK.

Q5i Analyzing the relationship between `percollege` and `percbelowpoverty` in `midwest`

We will now be analyzing the `midwest` dataset from `tidyverse`. We want to explore the relationship between `percollege` - the percentage of people who went to college in a county - and `percbelowpoverty` - the percent of that county living below the poverty line. We will be utilizing a scatterplot, as well as a smoothed line fit by the LOESS smoothing method. To reduce overplotting, we introduce an alpha of 0.5 to help see the concentration of data points.

```
ggplot(data = midwest, mapping = aes(x = percollege, y = percbelowpoverty)) +
  geom_point(alpha = 0.5) +
  geom_smooth() +
  labs(title = "Percentage of people below poverty line on
       Percentage of people who went to college within a county") +
  xlab("Percentage of people within the county who went to college") +
  ylab("Perecentage of people within the county who live
       below the poverty line")
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



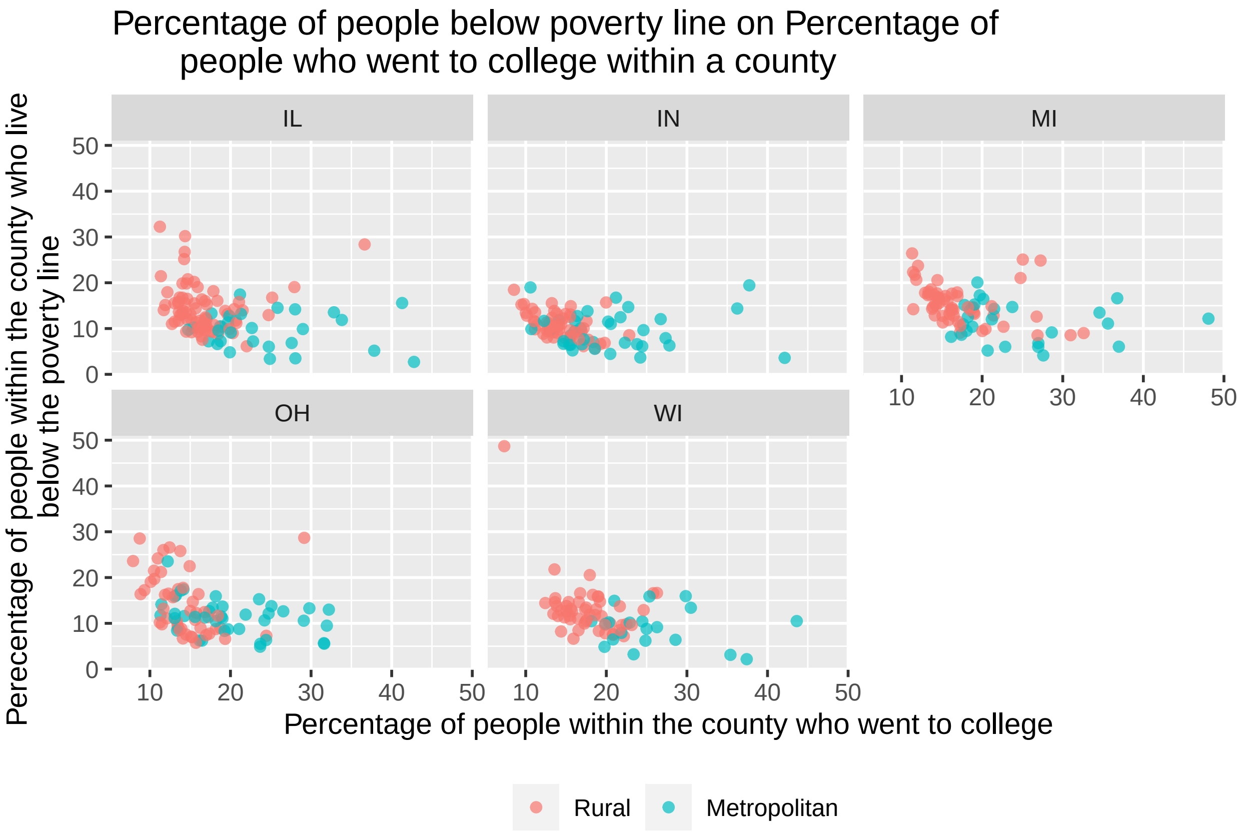
From this plot we can see an negative association between `percollege` and `percbelowpoverty` between 0 and around 15% of people within the county who went to college, and a fairly flat association for the rest of the domain. This means that for counties who had 0 to 15% of people go to college, we see higher percentages of people living below the poverty line in counties who had lower college attendance. There may or may not be a causal relationship but that's another question.

Q5ii Adding dimensions and beautifying the graph

We can add a lot more information to the graph we have above. Also, it could look nicer. We add the following information to the graph: whether a county is in an urban metropolitan area and the state that county is in. We also remove the smoothing lines to reduce clutter, increase the alpha to improve visibility, and switch up the colors.

```
ggplot(data = midwest,
       mapping = aes(x = percollege, y = percbelowpoverty,
                     color = as.factor(inmetro))) +
  geom_point(alpha = 0.7) +
  scale_color_brewer(palette = "Set1") +
  labs(title = "Percentage of people below poverty line on Percentage of
       people who went to college within a county") +
  xlab("Percentage of people within the county who went to college") +
  ylab("Perecentage of people within the county who live
       below the poverty line") +
  facet_wrap(~state) +
  theme(legend.position = "bottom", legend.title = element_blank()) +
  scale_color_discrete(labels = c("Rural", "Metropolitan"))
```

```
## Scale for colour is already present.
## Adding another scale for colour, which will replace the existing scale.
```



From this new plot we now have a lot more information, and can do some analysis. Firstly, we see that the blue (metropolitan) counties seem to be "lower" and more "to the right" than the red (rural) counties. This means that on average the metropolitan counties have a lower percentage of people living below the poverty line and a higher percentage of people who have gone to college. Regardless of if the county was in a metropolitan area, there does appear to be a general negative association between the percentage of people within the county who went to college and the percentage of people within the county who live below the poverty line. This seems to hold across different states as well.

There is an outlier in Jackson, IL. This state has 36.6% of people who have gone to college (high) but still 28.4% of people who live below the poverty line (high). This goes against the general trend we observed across the board.