Assignment2

2023-07-31

These exercises require you to generate plots of various kinds.

Part I – Still Visualization

These first few exercises will run through some of the simple principles of creating a ggplot2 object, assigning aesthetics mappings and geoms.

1. Used the flights data from the nycflights13 package.

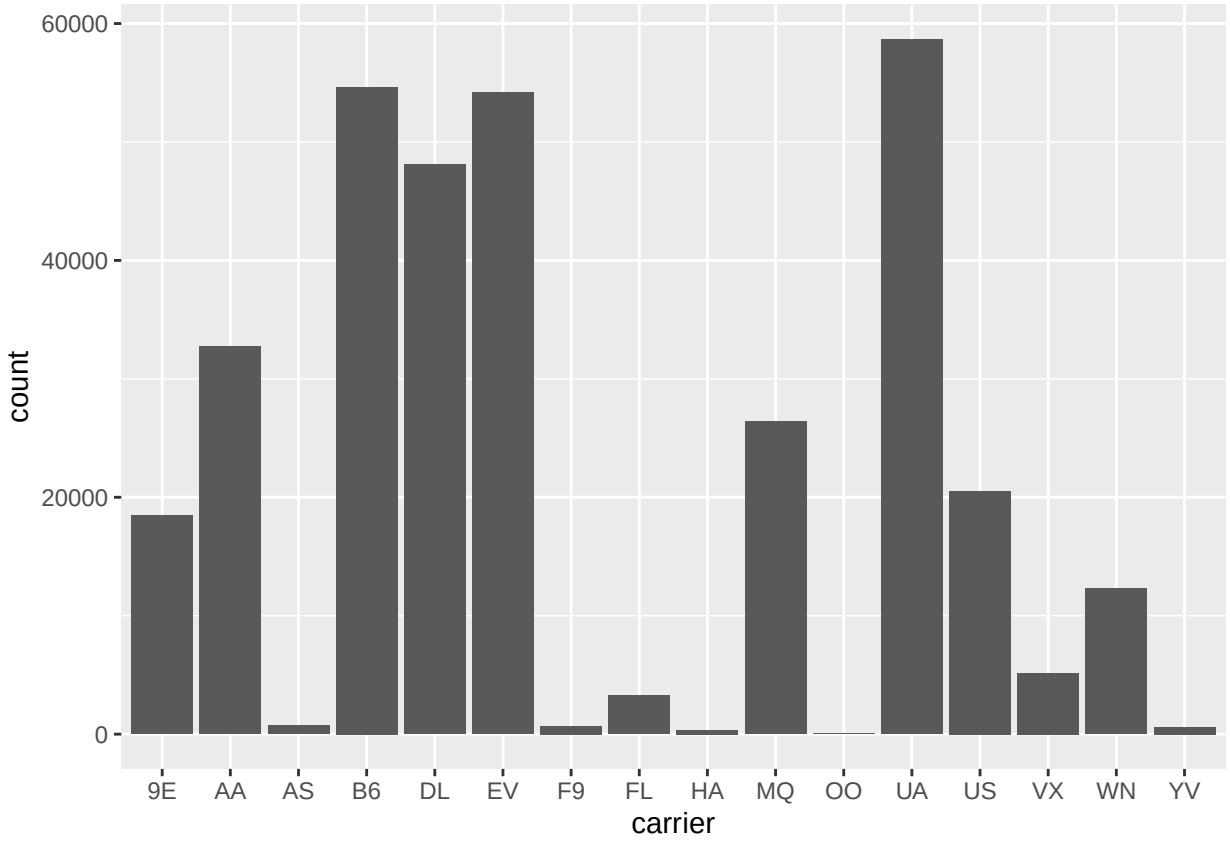
```
library(tidyverse)
library(nycflights13)
```

Bar plots with group by

2. Create a bar plot showing the number of flights flown out of New York airports by each carrier in 2013. Which airline carrier flew the most flights?

```
flights2 <- flights %>%
  ggplot(mapping=aes(carrier)) + geom_bar()

flights2
```

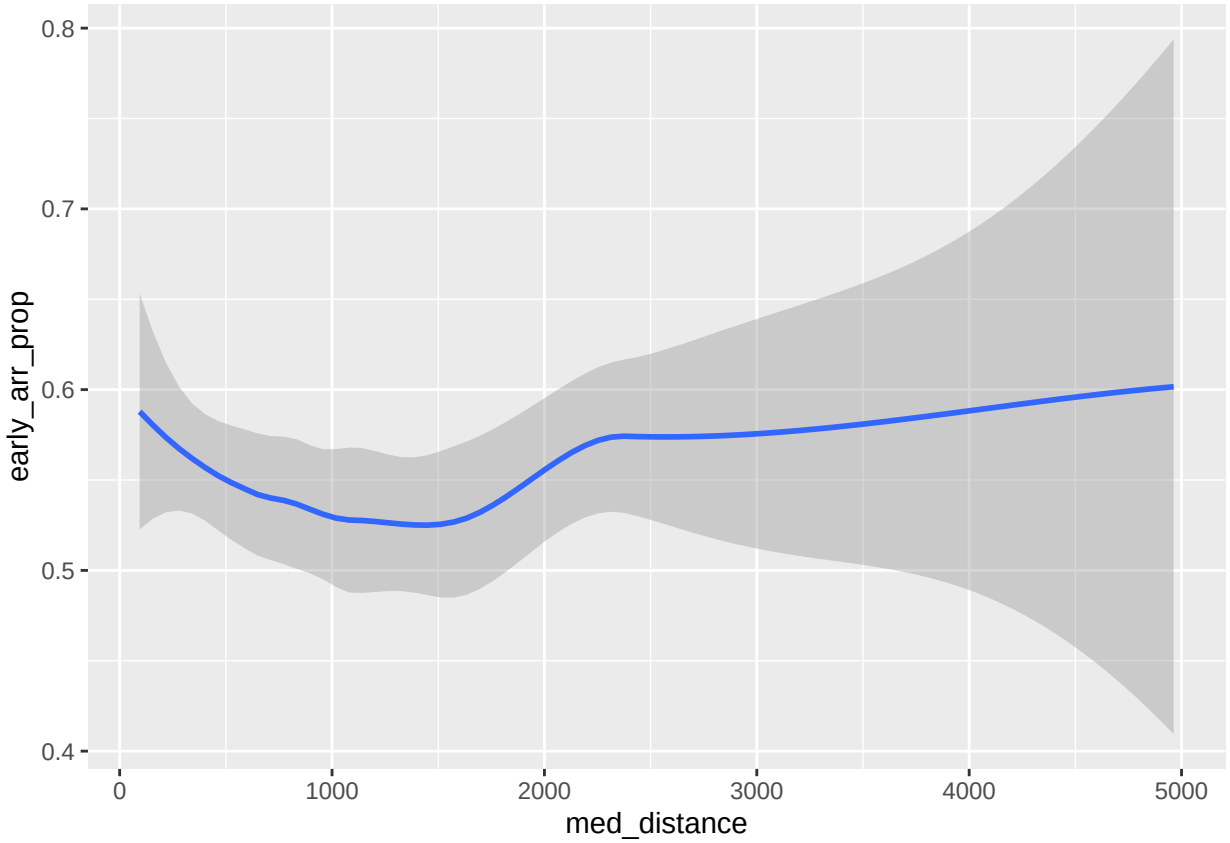



```
# Answer: UA flew the most flights in 2013
```

3. For each destination, calculate the proportion of flights that arrived at their destination earlier than scheduled. Also calculate the median distance flown to each destination. Plot the proportion of early arrivals (on the y-axis) against the median distance flown (on the x-axis) for each destination. Describe the relationship between early arrivals and flight distance if you see any patterns [Hint: geom_smooth() is a function we didn't talk about in the class, but it might help you a little]

```
flights3 <- flights %>%
  drop_na() %>%
  group_by(dest) %>%
  summarize(
    early_arr_prop = mean(arr_delay < 0),
    med_distance = median(distance)) %>%
  ggplot(mapping=aes(x=med_distance, y=early_arr_prop)) +
  geom_smooth()

flights3
```




```
# Answer: proportion of early arrival is dropping from 0.6 to around 0.53 when median distance reaches
```

Variable transformation

4. Read in the healthcare-dataset-stroke-data ,into a new object called 'stroke' with function read.csv

```
stroke = read_csv("healthcare-dataset-stroke-data.csv")
```

```
## Rows: 5110 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (6): gender, ever_married, work_type, Residence_type, bmi, smoking_status
## dbl (6): id, age, hypertension, heart_disease, avg_glucose_level, stroke
##
## 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.
```

```
head(stroke)
```

```
## # A tibble: 6 x 12
##      id gender   age hypertension heart_disease ever_married work_type
##   <dbl> <chr>  <dbl>        <dbl>         <dbl> <chr>        <chr>
## 1  9046 Male      67            0             1 Yes          Private
```

```
## 2 51676 Female    61            0             0 Yes          Self-employed
## 3 31112 Male      80            0             1 Yes          Private
## 4 60182 Female    49            0             0 Yes          Private
## 5  1665 Female    79            1             0 Yes          Self-employed
## 6 56669 Male      81            0             0 Yes          Private
## # i 5 more variables: Residence_type <chr>, avg_glucose_level <dbl>, bmi <chr>,
## #   smoking_status <chr>, stroke <dbl>
```

```
view(stroke)
```

5a. Subset the male patient and create two more columns, one is the square of bmi, another is the square of avg_glucose_level

```
stroke5a <- stroke %>%
  filter(gender == "Male") %>%
  mutate(bmi_sqrt = as.numeric(bmi)^2,
         avg_glucose_level_sqrt = avg_glucose_level^2)
```

```
## Warning: There was 1 warning in `mutate()`.
## i In argument: `bmi_sqrt = as.numeric(bmi)^2`.
## Caused by warning:
## ! NAs introduced by coercion
```

```
# Answer is in "stroke5a"
view(stroke5a)
```

5b. Subset the female patient and also find the patient that bmi is between 15 to 25.

```
stroke5b <- stroke %>%
  filter(gender == "Female", between(as.numeric(.$bmi), 15, 25))
```

```
## Warning: There was 1 warning in `filter()`.
## i In argument: `between(as.numeric(.$bmi), 15, 25)`.
## Caused by warning in `between()`:
## ! NAs introduced by coercion
```

```
# Answer is in "stroke5b"
view(stroke5b)
```

6. For the male population, find the proportion of people that has stroke

```
stroke6 <- stroke %>%
  filter(gender == "Male") %>%
  summarize(stroke_proportion = mean(.$stroke == 1))

stroke6
```

```
## # A tibble: 1 x 1
##   stroke_proportion
##               <dbl>
## 1            0.0511
```

7. For each gender, find the proportion of people that has stroke

```
stroke7 <- summarize(group_by(stroke, gender), gender_proportion = mean(stroke == 1))

stroke7
```

```
## # A tibble: 3 x 2
##   gender gender_proportion
##   <chr>              <dbl>
## 1 Female            0.0471
## 2 Male              0.0511
## 3 Other             0
```

8. Find the number of people for each smoking status.

```
stroke8 <- summarize(group_by(stroke, smoking_status), smoking_status_population = n())

stroke8
```

```
## # A tibble: 4 x 2
##   smoking_status  smoking_status_population
##   <chr>                              <int>
## 1 Unknown                             1544
## 2 formerly smoked                      885
## 3 never smoked                        1892
## 4 smokes                               789
```