assingment 5 redo

2024-09-29

```
#install.packages("dplyr")
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## 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 become errors
```

```
library(readr)
library(tidyr)

library(dplyr)
data <- read.table(file = "C:/Users/dbrusche/Desktop/wide_airport.csv", header = TRUE, sep = "\t")
```

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
# renaming columns
colnames(data)[1] <- "Airline"
colnames(data)[2] <- "Arrival"
data[2, 1] <- "ALASKA"
data[5, 1] <- "AM WEST"

# Transform to long format and remove NAs
long_data <- data %>%
  pivot_longer(
    cols = starts_with("Los"):Seattle,
    names_to = "Destination",
    values_to = "Frequency"
  ) %>%
  drop_na()  # This will remove rows with NA in the Delay column
```

Including Plots

You can also embed plots, for example:

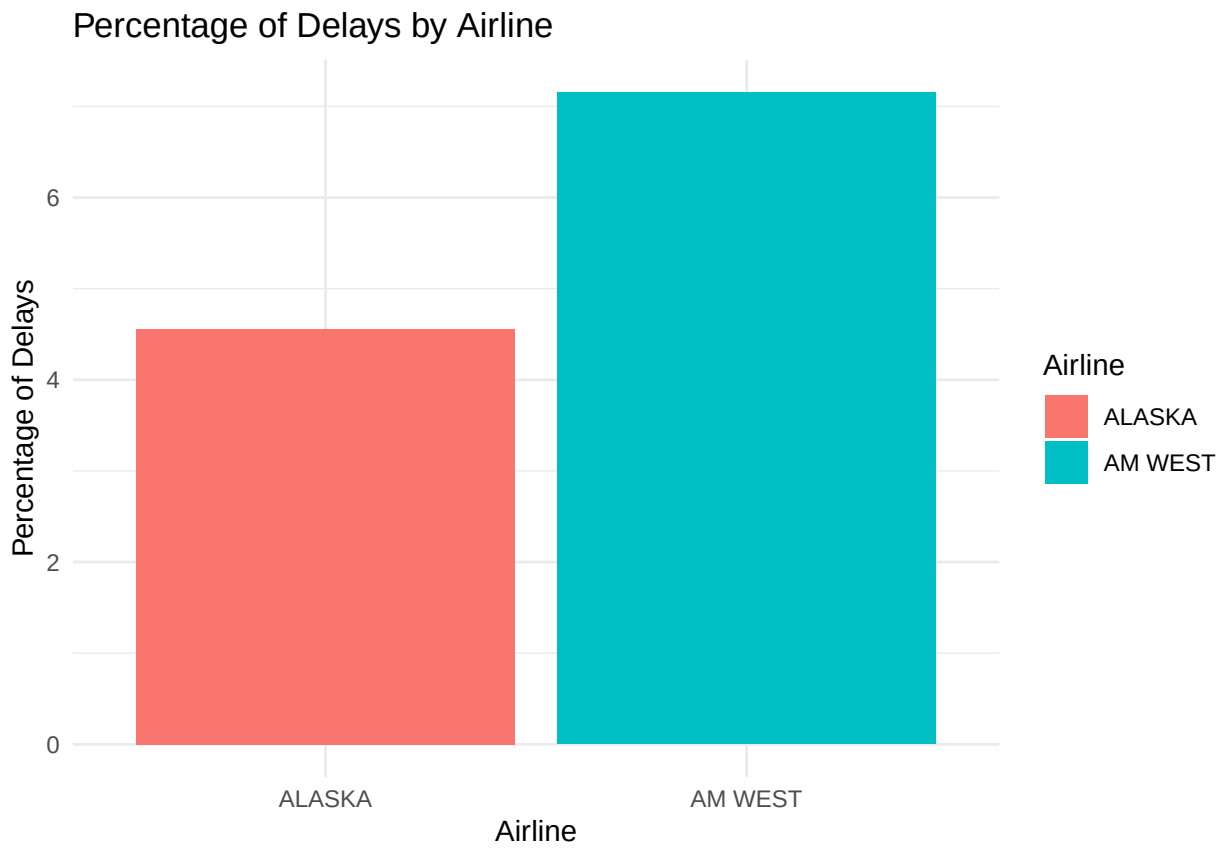
```
summary_data <- long_data %>%
  group_by(Airline, Arrival) %>%
  summarise(Total = sum(Frequency), .groups = 'drop') %>%
  mutate(Percentage = Total / sum(Total) * 100)

# Filter only delays
delay_data <- summary_data %>%
  filter(Arrival == "delayed")

# Print the summary table
print("Summary of Delays:")
```

```
## [1] "Summary of Delays:"
```

```
print(delay_data)
```

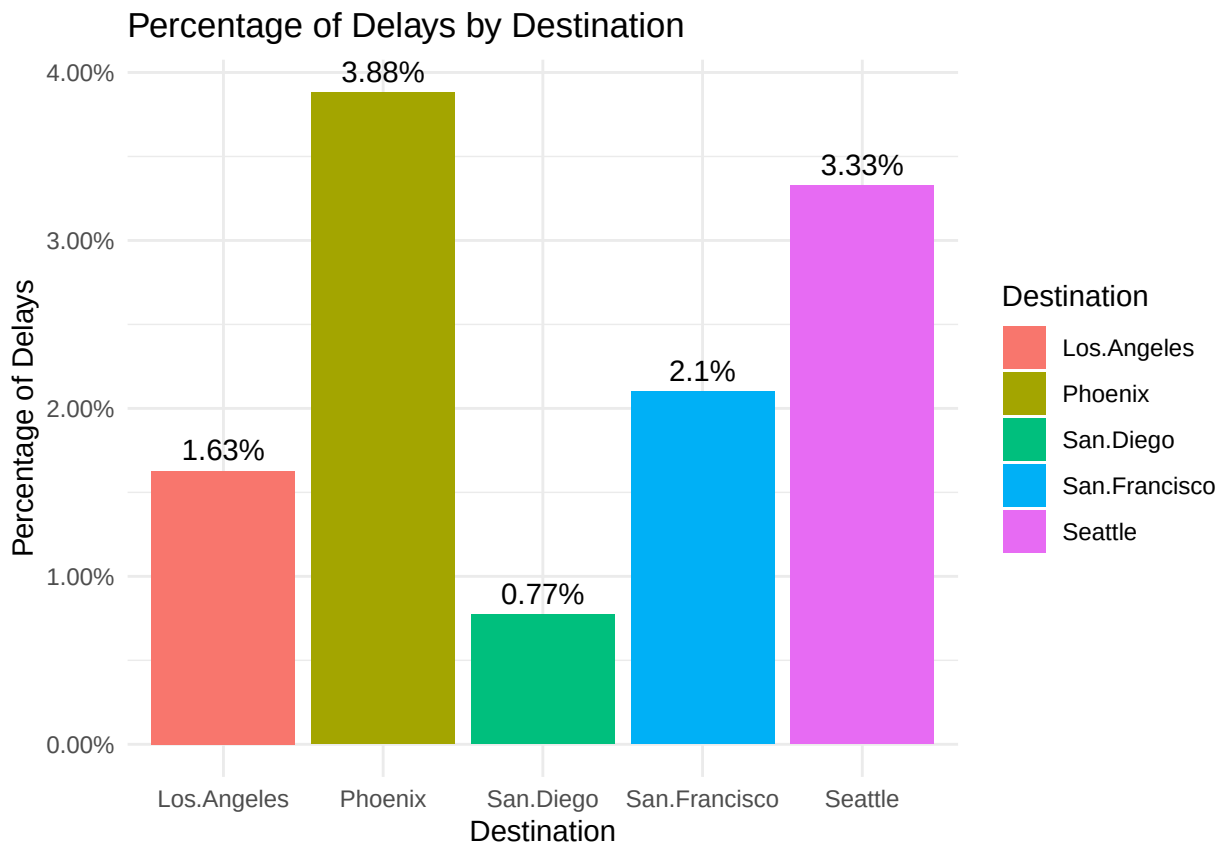
```
## # A tibble: 2 x 4
##   Airline Arrival Total Percentage
##   <chr>   <chr>   <int>      <dbl>
## 1 ALASKA  delayed   501       4.55
## 2 AM WEST delayed   787       7.15
```

```
##barplot
ggplot(delay_data, aes(x = Airline, y = Percentage, fill = Airline)) +
  geom_bar(stat = "identity") +
  labs(title = "Percentage of Delays by Airline",
       x = "Airline",
       y = "Percentage of Delays") +
  theme_minimal()
```




```
#From the summary data, we see that the total count for Alaska Airlines was 501 delayed flights and 3,2
```

```
summary_data2 <- long_data %>%
  group_by(Destination, Arrival) %>%
  summarise(Total = sum(Frequency), .groups = 'drop') %>%
  mutate(Percentage = Total / sum(Total) * 100)

delay_data2 <- summary_data2 %>%
  filter(Arrival == "delayed")


ggplot(delay_data2, aes(x = Destination, y = Percentage, fill = Destination)) +
  geom_bar(stat = "identity") +
  labs(title = "Percentage of Delays by Destination",
       x = "Destination",
       y = "Percentage of Delays") +
  theme_minimal() +
  scale_y_continuous(labels = scales::percent_format(scale = 1, accuracy = 0.01)) +
  geom_text(aes(label = paste0(round(Percentage, 2), "%")), vjust = -0.5)
```




```
#From the summary, we can see the delayed and on-time counts for each destination along with their perc

#The discrepancy in examining only the airlines for arrival data is the lack of information it provides
```