Project 3

Brian Folkers (bdf676)

This is the dataset used in this project:

```r
transit_cost <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/da

# backup dataset: https://github.com/rfordatascience/tidytuesday/blob/master/data/2021/2021-06-22/readm

transit_cost <- na.omit(transit_cost) #remove na rows
transit_cost
```

```
## # A tibble: 428 x 20
##        e country city  line  start_year end_year    rr length tunnel_per tunnel
##    <dbl> <chr>   <chr> <chr> <chr>      <chr>    <dbl>  <dbl> <chr>       <dbl>
##  1  7136 CA      Vanc~ Broa~ 2020       2025         0    5.7 87.72%        5
##  2  7137 CA      Toro~ Vaug~ 2009       2017         0    8.6 100.00%       8.6
##  3  7138 CA      Toro~ Scar~ 2020       2030         0    7.8 100.00%       7.8
##  4  7139 CA      Toro~ Onta~ 2020       2030         0   15.5 57.00%        8.8
##  5  7144 CA      Toro~ Yong~ 2020       2030         0    7.4 100.00%       7.4
##  6  7145 NL      Amst~ Nort~ 2003       2018         0    9.7 73.00%        7.1
##  7  7146 CA      Mont~ Blue~ 2020       2026         0    5.8 100.00%       5.8
##  8  7147 US      Seat~ U-Li~ 2009       2016         0    5.1 100.00%       5.1
##  9  7152 US      Los ~ Purp~ 2020       2027         0    4.2 100.00%       4.2
## 10  7153 US      Los ~ Purp~ 2018       2026         0    4.2 100.00%       4.2
## # ... with 418 more rows, and 10 more variables: stations <dbl>, source1 <chr>,
## #   cost <dbl>, currency <chr>, year <dbl>, ppp_rate <dbl>, real_cost <chr>,
## #   cost_km_millions <dbl>, source2 <chr>, reference <chr>
```

Link to the dataset: *https://github.com/rfordatascience/tidytuesday/blob/master/data/2021/2021-01-05/readme.md*

Part 1

Question: Is there a linear relationship between between the number of stations and the cost for completed lines in China, India, and Japan?

Introduction: The `transit_cost` dataset contains information on international public transit project including the details of its construction and cost. The main purpose of the dataset is to investigate what varaibles affect the cost of the projects and why it varies by region. For Part 1, the objective will be to examine how the number of `stations` in a line affects the cost per kilometer (`cost_km_millions`). In addition, this analysis will focus on the countries `China`, `Japan`, and `India`. These three countries have the most completed lines in the dataset and have different economies that will likely add to variation in the data.

Approach: The data wrangling for this section will simply focus on limiting the data to transit projects that are completed and are in one of the three aforementioned countries. The analysis will use a linear regression analysis. This will allow for a basic view of the relationship between stations and cost and if the two correlate with eachother for these countries. The analysis will create both a summary table of the linear regressions and a graph.

Analysis:

```
#needed libraries
library(broom) # for making linear regression summary table
library(glue)  # for easy text formatting
```

```
##
## Attaching package: 'glue'

## The following object is masked from 'package:dplyr':
##
##     collapse
```

```
#wrangling
transit_lm <- transit_cost %>% #new df
  filter(tunnel_per != "100.00%") %>% #filter for only lines that are completed
  mutate(country = replace(country, country == "CN", "China")) %>% #renaming
  mutate(country = replace(country, country == "IN", "India")) %>% #renaming
  mutate(country = replace(country, country == "JP", "Japan")) %>% #renaming
  filter(country == "China" | country == "India" | country == "Japan") #filtering for three countries

#summary table of linear regression for each country
lm_summary <- transit_lm %>%
  nest(data = -country) %>%
  mutate( # apply linear model to each nested data frame
    fit = map(data, ~lm(cost_km_millions ~ stations, data = .x)),
    glance_out = map(fit, glance)
  ) %>%
  select(country, glance_out) %>%
  unnest(cols = glance_out)
lm_summary
```

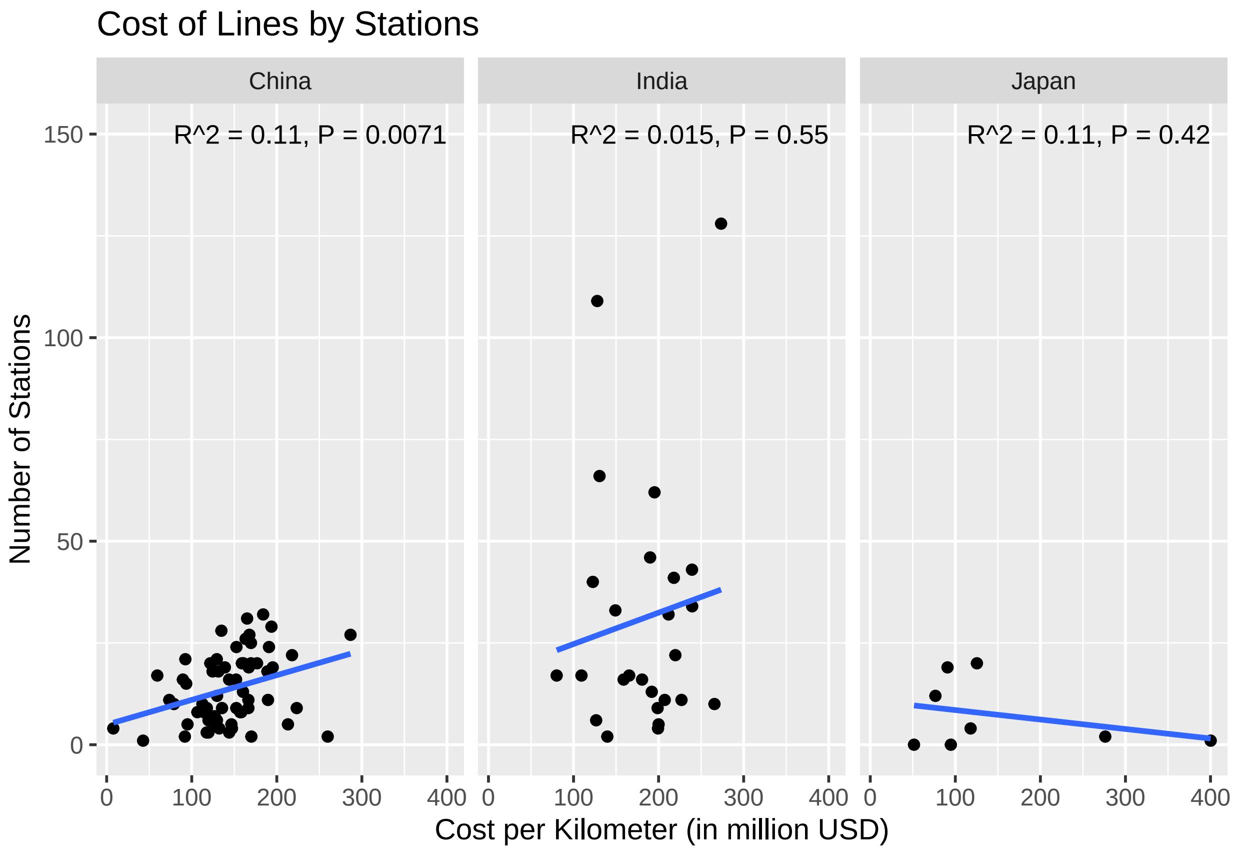
```
## # A tibble: 3 x 13
##   country r.squared adj.r.squared sigma statistic p.value    df logLik   AIC
##   <chr>       <dbl>         <dbl> <dbl>     <dbl>   <dbl> <dbl>  <dbl> <dbl>
## 1 India      0.0149       -0.0262  50.0     0.363 0.553       1 -138.   281.
## 2 Japan      0.109        -0.0397 123.      0.733 0.425       1  -48.7  103.
## 3 China      0.115         0.100   45.0     7.78  0.00707     1 -323.   652.
## # ... with 4 more variables: BIC <dbl>, deviance <dbl>, df.residual <int>,
## #   nobs <int>
```

```
#make label data to put on plot
label_data <- lm_summary %>%
  mutate(
    rsqr = signif(r.squared, 2), #round to 2 significant digits
    pval = signif(p.value, 2),
    label = glue("R^2 = {rsqr}, P = {pval}"),
    cost_km_millions = 400, stations = 150 #label position in plot
  ) %>%
  select(country, label, cost_km_millions, stations)
label_data
```

```
## # A tibble: 3 x 4
##   country label                     cost_km_millions stations
##   <chr>   <glue>                               <dbl>    <dbl>
## 1 India   R^2 = 0.015, P = 0.55                  400      150
## 2 Japan   R^2 = 0.11, P = 0.42                   400      150
## 3 China   R^2 = 0.11, P = 0.0071              400      150
```

```
#make linear regression plot
ggplot(transit_lm, aes(cost_km_millions, stations)) + geom_point() +
  geom_text(
    data = label_data, aes(label = label), #add label data
    size = 10/.pt, hjust = 1  # 10pt, right-justified
  ) +
  geom_smooth(method = "lm", se = FALSE) + # add linear regression line
  facet_wrap(vars(country))  + #facet by country
  xlab("Cost per Kilometer (in million USD)") + #labeling
  ylab("Number of Stations") + #labeling
  ggtitle("Cost of Lines by Stations") #labeling
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Discussion: The results of this analysis reveal a few interesting trends. Firstly, only China has a significant correlation (p = 0.0071). Both China and India have positive correlations, indicating that more stations tends to increase cost, while Japan has a negative correlation (though this may just be the result of a lack of datapoints). In addition, all models have very low r-squared, which would suggest that they explain very little of the variation in the data. Though this linear regression provides a tangible start for an analysis of this data, it is clear that more data on completed rail lines is needed before any conclusive statements can be made on how stations affect transit constriction cost.

Part 2

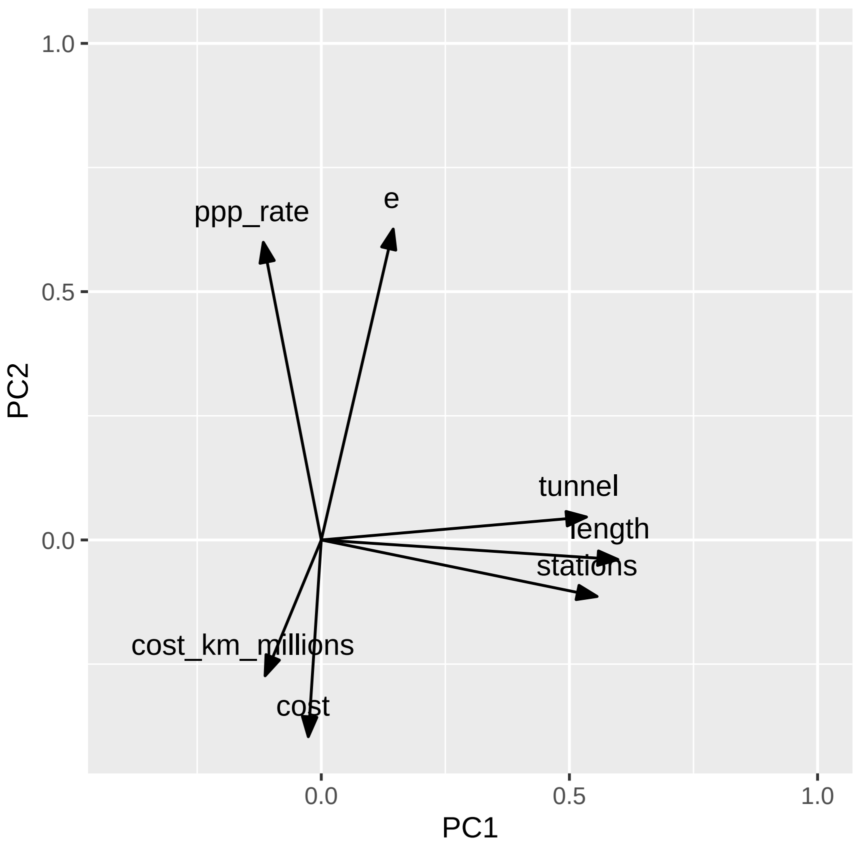
Question: How are all of the numeric variables in the `transit_cost` dataset correlated?

Introduction: For Part 2, the objective will be to examine how the different numeric variables correlate. Therefore it will mainly focus on the rail road status as the separating categorical variable, and the numeric variables like `length`, `stations`, `cost`, and `ppp_rate`.

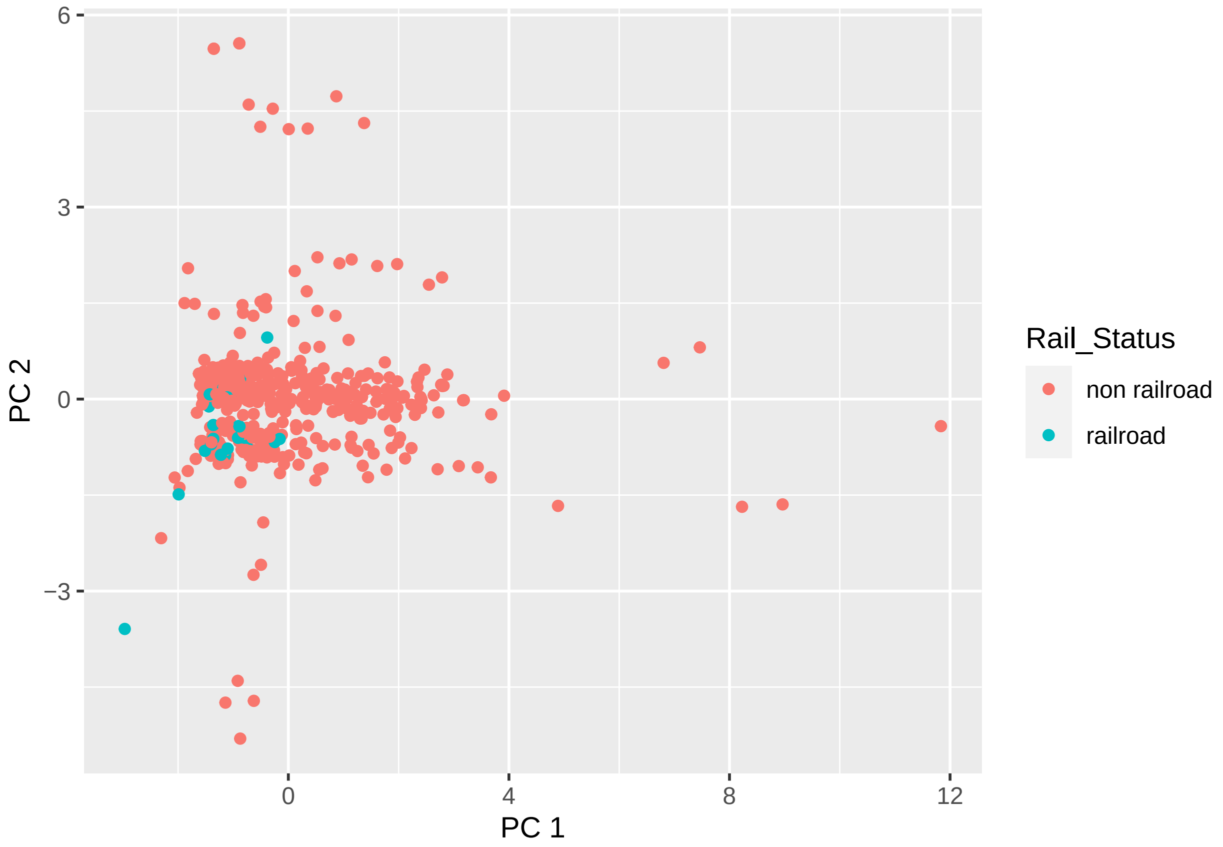
Approach: The analysis will use a Principle Component Analysis (PCA). This will allow for the analysis of the correlation all numeric factors on a single graph. The data wrangling will mostly focus on renaming variables and removing some of the year variables that cause errors in the code. The PCA will consist of a rotation matrix and a PCA scatter plot, which will be used to determine what variables affect the cost_km_millions. It will also include a variance explained bar graph. This section will include all transit projects, even if they have not finished their construction.

Analysis:

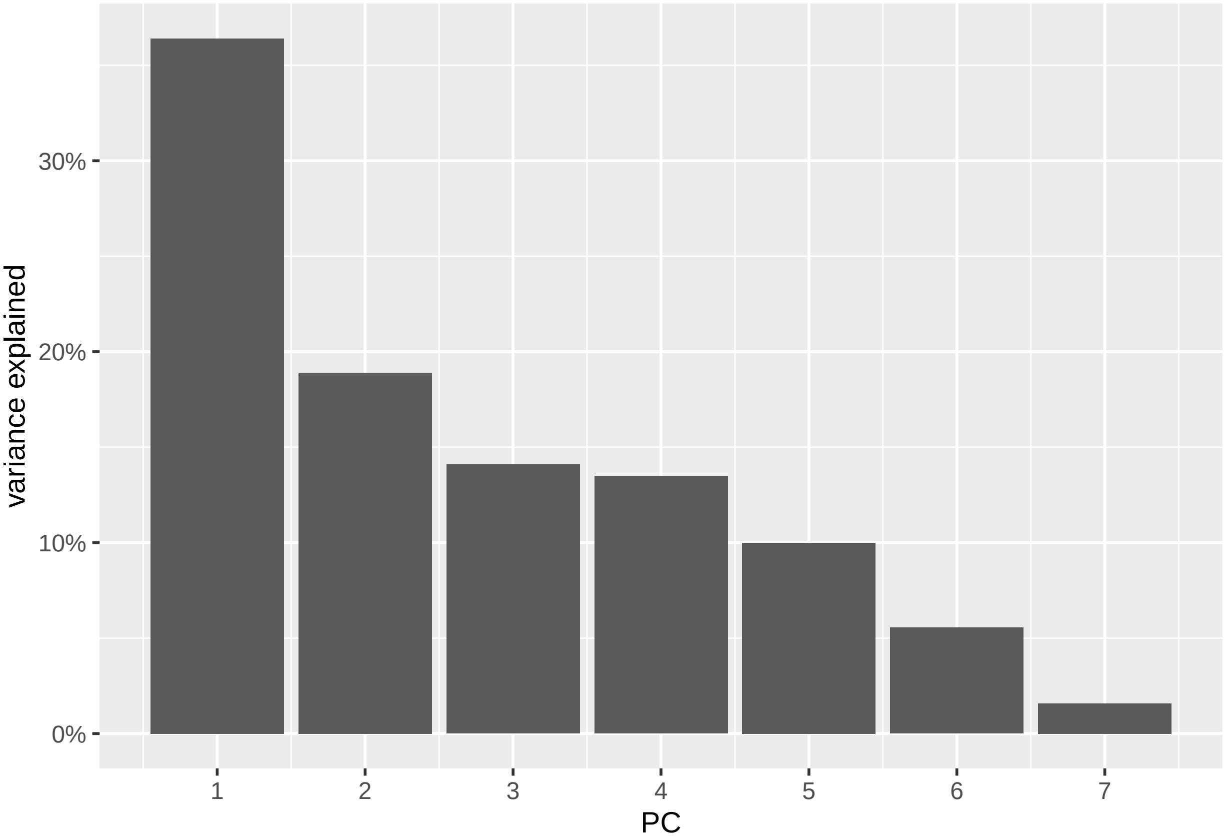
```r
#wrangling
transit_pca <- transit_cost %>%
  na.omit() %>%
  mutate(rr = replace(rr, rr == 0, "non railroad")) %>% #renaming
  mutate(rr = replace(rr, rr == 1, "railroad")) %>% #renaming
  select(-c('end_year', 'start_year', 'year')) %>% #removing
  rename(Rail_Status = rr) %>% #renaming
  na.omit() #remove NAs


#First make a PCA with prcomp()
pca_fit <- transit_pca %>%
  select(where(is.numeric)) %>% # retain only numeric columns
  scale() %>%                   # scale to zero mean and unit variance
  prcomp()                     # do PCA


#rotation matrix plot
arrow_style <- arrow(
  angle = 20, length = grid::unit(8, "pt"),
  ends = "first", type = "closed"
)
pca_fit %>%
  # extract rotation matrix
  tidy(matrix = "rotation") %>%
  pivot_wider(
    names_from = "PC", values_from = "value",
    names_prefix = "PC"
  ) %>%
  ggplot(aes(PC1, PC2)) +
  geom_segment(
    xend = 0, yend = 0,
    arrow = arrow_style
  ) +
  geom_text(aes(label = column), hjust = 0.6, vjust = -1) + #adjust labels
  xlim(-0.4, 1) + ylim(-0.4, 1) + #change axes' ranges
  coord_fixed()
```



```
#PCA scatterplot
library(broom)  # for augment(), tidy()
pca_fit %>%
  # add PCs to the original dataset
  augment(transit_pca) %>%
  ggplot(aes(.fittedPC1, .fittedPC2)) +
  xlab("PC 1") +
  ylab("PC 2") +
  geom_point(aes(color = Rail_Status)) +
  labs(fill = "Rail Status") #this wouldn't work for some reason
```



```
#variance explained plot
pca_fit %>%
  # extract eigenvalues
  tidy(matrix = "eigenvalues") %>%
  ggplot(aes(PC, percent)) +
  geom_col() +
  scale_x_continuous(
    # create one axis tick per PC
    breaks = 1:8 #8 shown in rotation matrix
  ) +
  scale_y_continuous(
    name = "variance explained",
    # format y axis ticks as percent values
    label = scales::label_percent(accuracy = 1)
  )
```



Discussion: Judging by the PCA plot, railroads and non-railroads separate mostly along PC1. Looking at the rotation matrix, this would indicate that transit projects have a higher `cost_km_millions` with more length, stations, and tunnels. It is also interesting to note that purchasing power parity (`ppp_rate`) does not have much affect on the cost. Finally, the variance explained chart would suggest that both PC 1 and 2 explain a fairly low amount of the variance in the data (about 25% and 20% respectively). It is interesting to note that the main physical attributes of the projects appear to have the most effect on its cost.