Multivariate Data Analysis: Airline Data

Huong Ngo

3/17/2022

Outline

- What is Multivariate Data Analysis?
- The Dataset
- Preliminary Data Visualizations
- Principal Components Analysis (PCA)
- K-Means Clustering
- Introducing Other Methods for Future Work
- What I Learned!

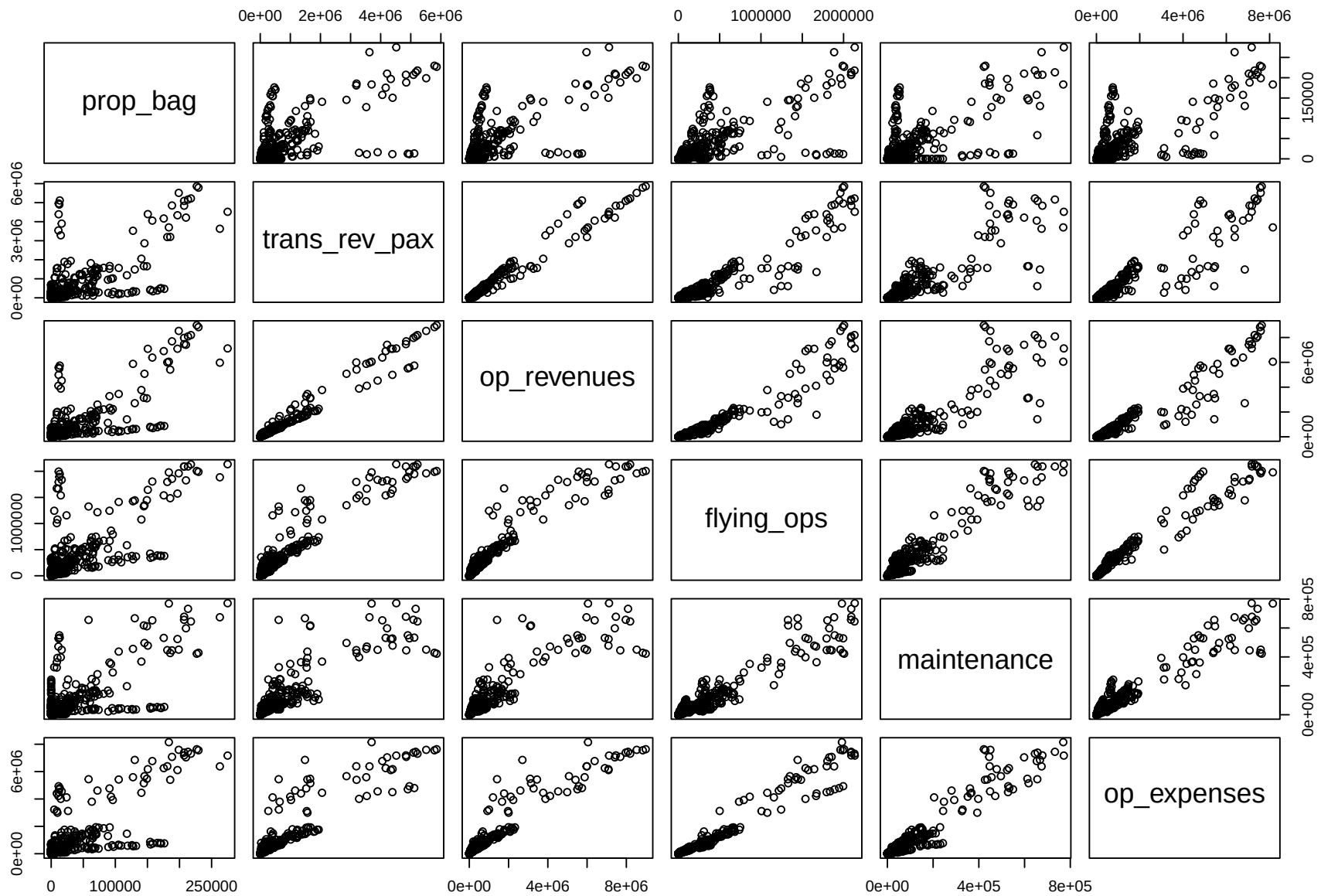
What is Multivariate Data Analysis

- From *STAT 311*, students learned about **univariate data analysis**: analysis on data that is of **one (univariate)** variable
- Multivariate data: data that **contains values recorded for** *multiple different* **variables**
- Multivariate data analysis: *simultaneous* **statistical analysis of a collection of variables**
- Main goal: **uncover, display or extract any "signal" in the data in the presence of noise to discover what the data has to tell us**

The Dataset

- Examined operations financial data of passenger airline carriers from 2019 - 2021
- 26 variables
  - Revenue sources: *Baggage Fees*, *Cancellation Fees*, *Transportation Services*
  - Expense sources: *Maintenance*, *Aircraft Services*, *Passenger Services*
  - Overall information: *Net Income*, *Operation Profit/Loss*, *Operation Revenues/Expenses*
- 496 observations
  - Each observation is the financial details of an airline carrier in a quarter of a specific year in a specific region
- Excluded cargo and charter airliners

Preliminary Data Visualizations



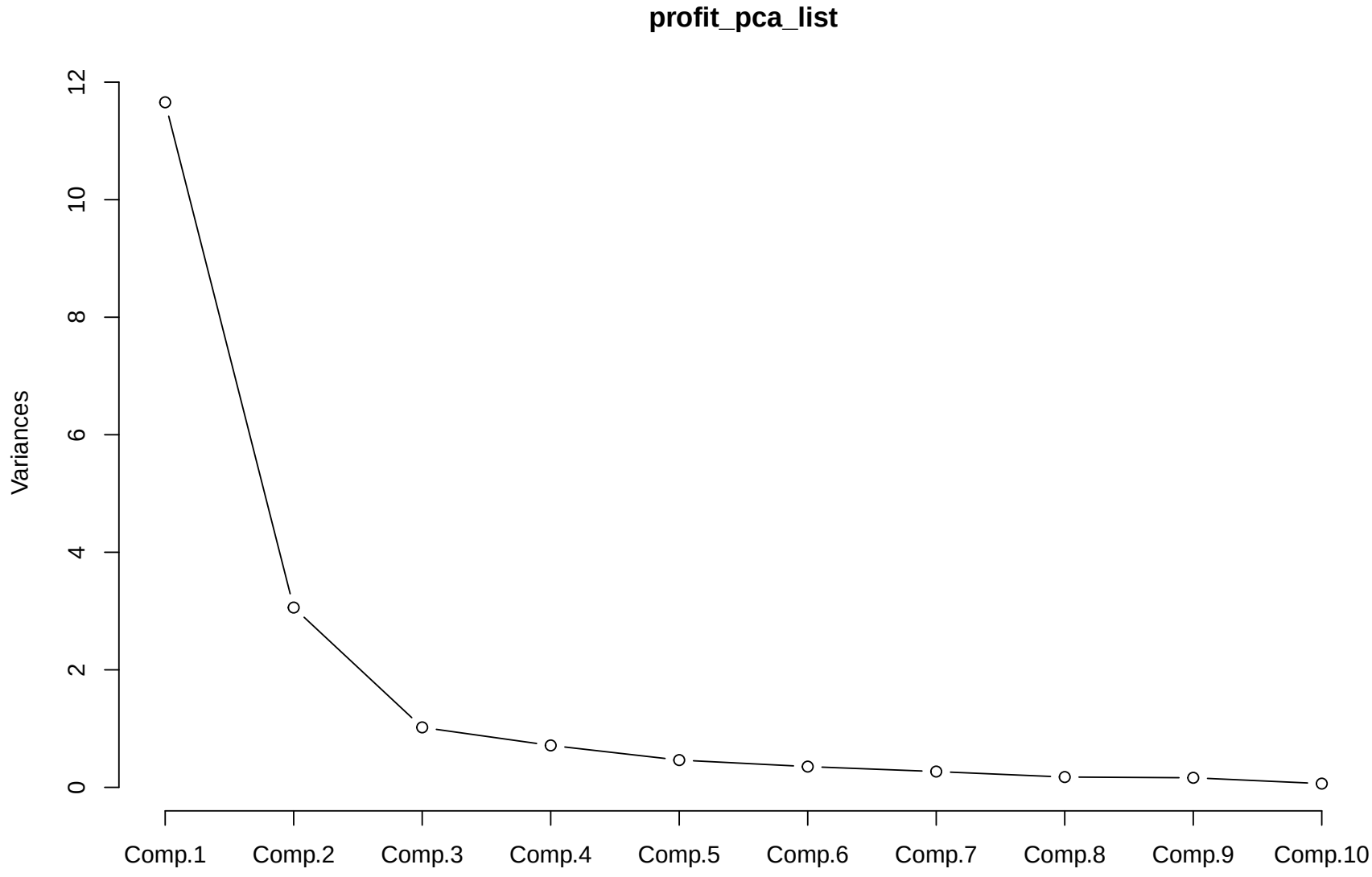
Motivations from Data Visualizations and Nature of Dataset

- Too many variables!
- Correlation between variables in dataset: Collinearity!
- Solution:
  - **PCA**: Reduce dimensionality of dataset while retaining as much variation of the original dataset
- Clusters in Scatterplot Matrix
  - **K-Means Clustering**: Cluster analysis that aims to uncover clusters of observations that are homogeneous

Principal Components Analysis (PCA)

- *princomp()*: PCA function that performs a series of calculations to describe variation of a set of correlated variables as a new, uncorrelated set of variables.
  - New set of variables are the linear combinations of the old variables
  - We will use the first few variables (2 or 3) as they will capture the most variation and give a lower-dimensional summary of data. These are called the principal components
- Standardizing dataset: The variables have different range of values
  - To ensure our analysis is not affected by the scales, we will standardize it

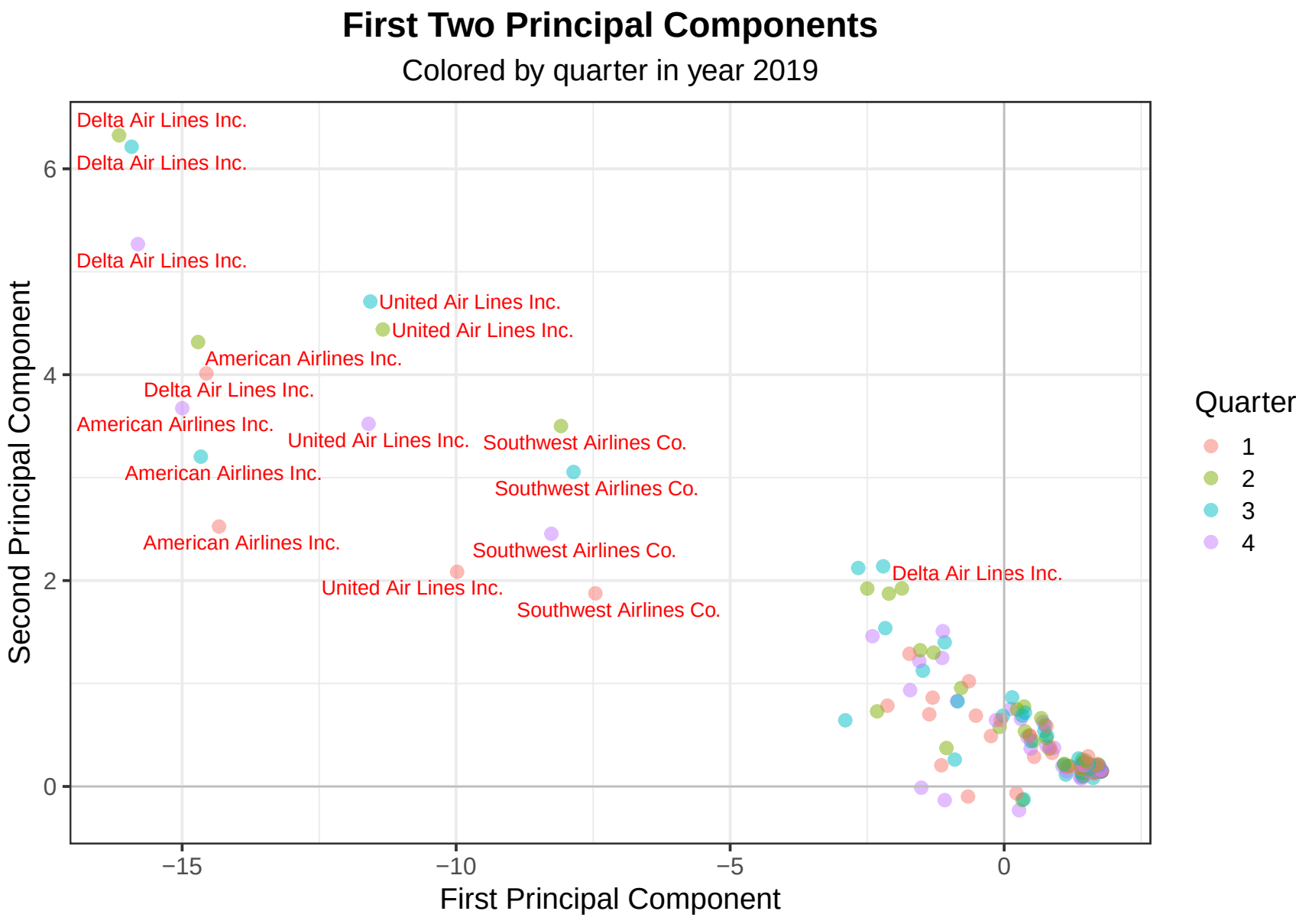
Principal Components Analysis (PCA)



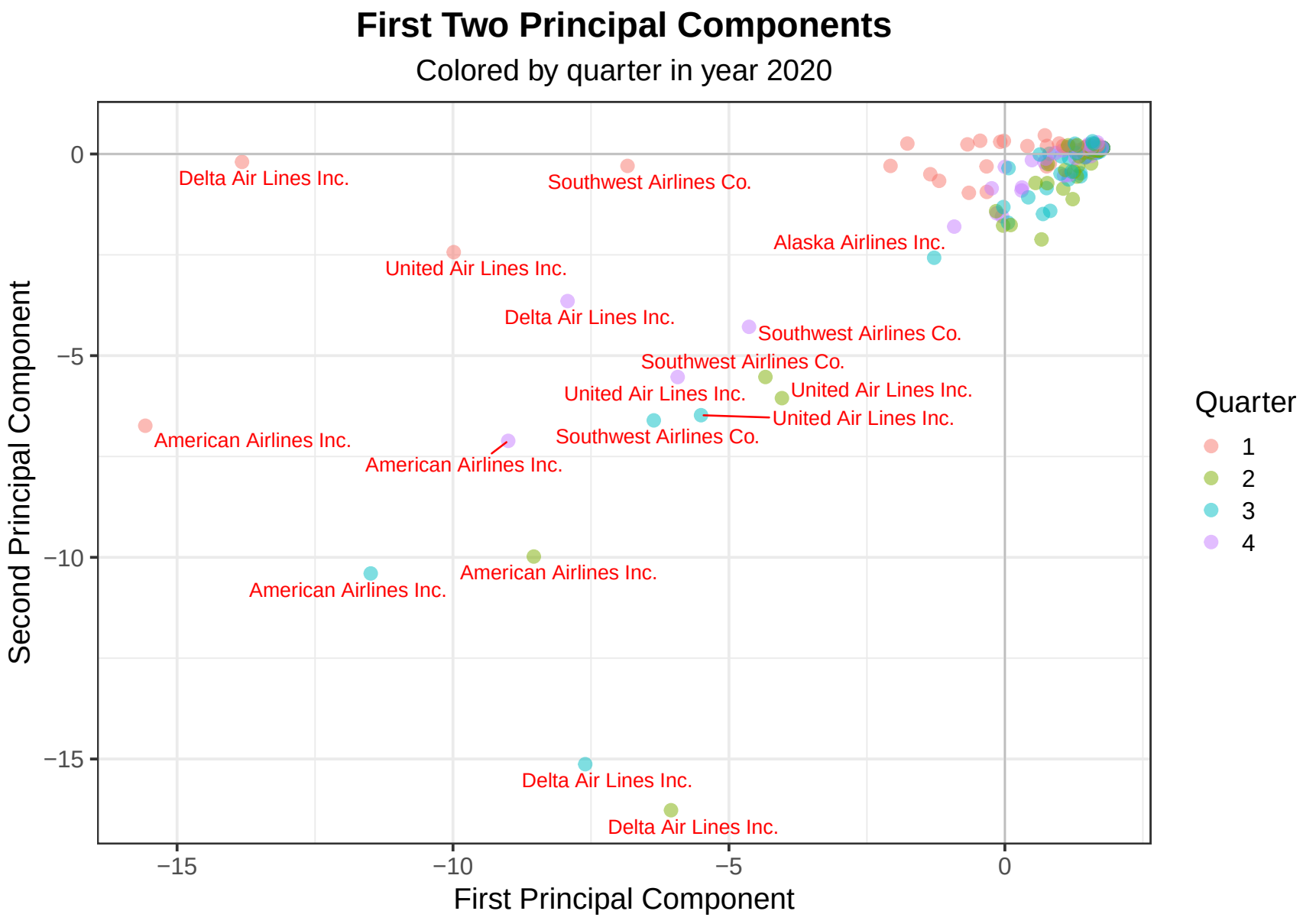
Interpretation of Principal Components Analysis (PCA)

- First Principal Component is characterized by **revenue and expense sources**
- Second Principal Component is characterized by **general financial information** about a carrier
- Third Principal Component is characterized by **revenue gained from moving freight**

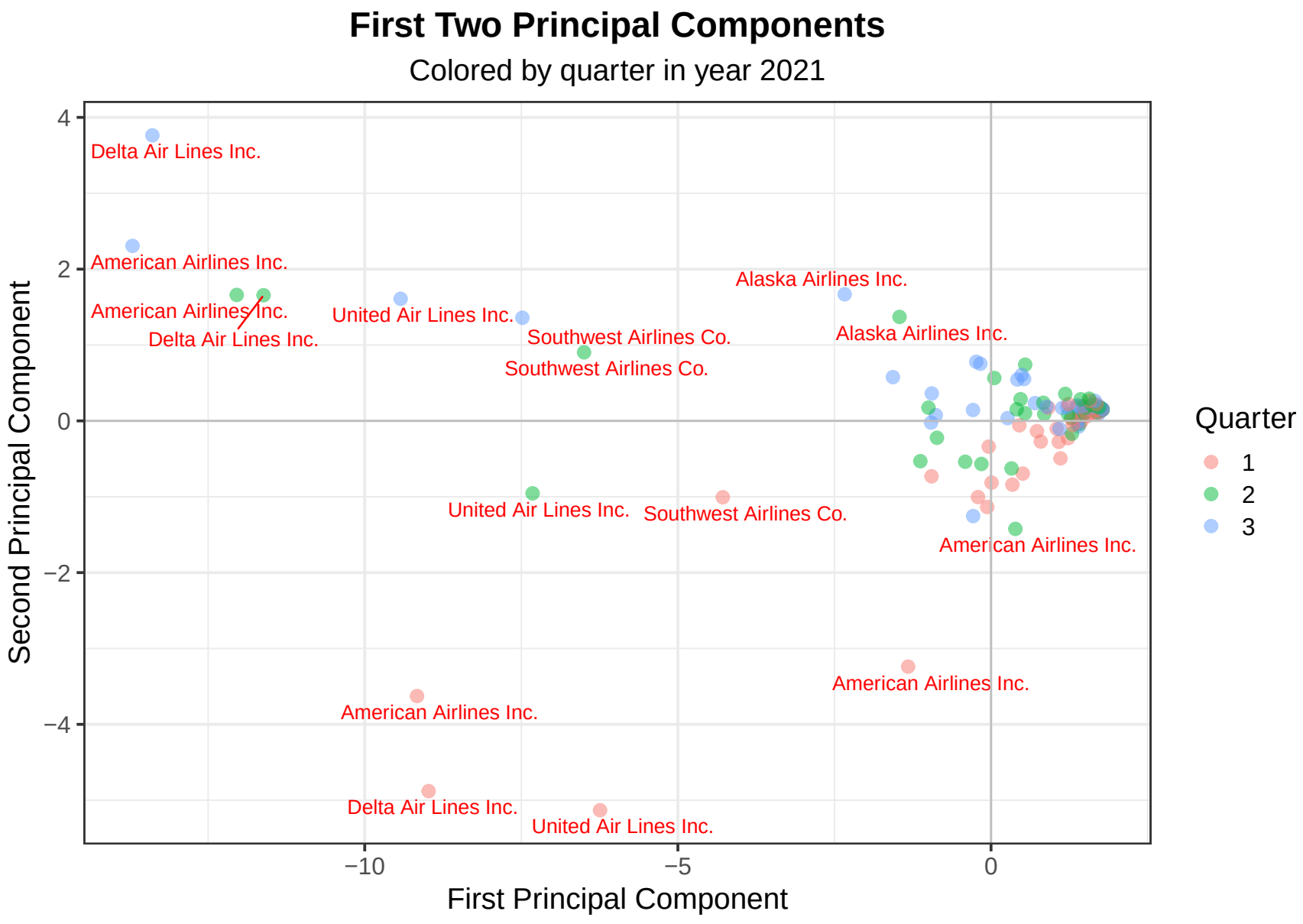
Principal Components Analysis (PCA)



Principal Components Analysis (PCA)



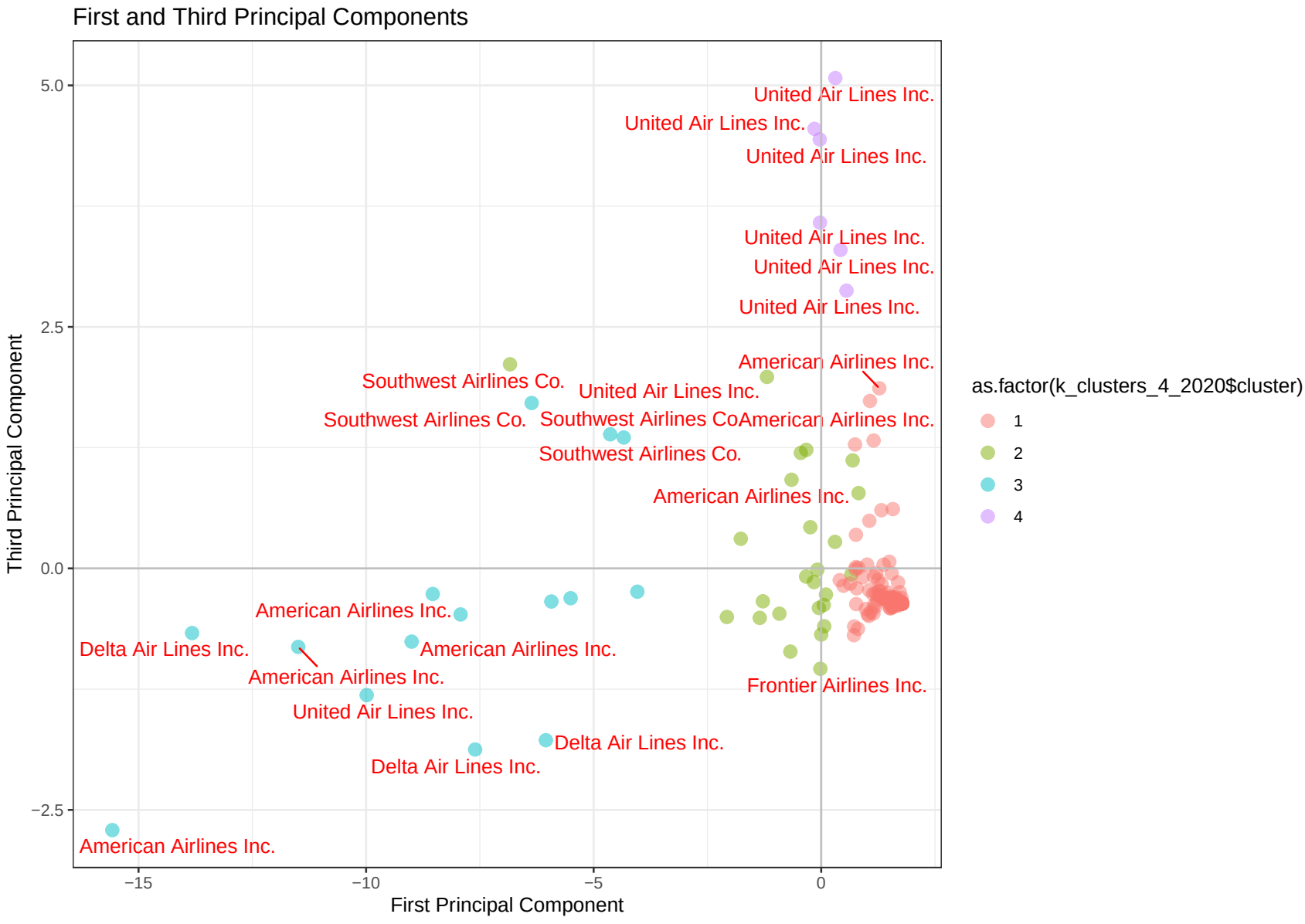
Principal Components Analysis (PCA)



K-Means Clustering

- **kmeans()** - K-Means Clustering algorithms function that partitions n observations in a dataset into k clusters that minimizes a numerical criterion
  - Within-Group Sum of Squares (WGSS)
  - Intuitively, it is a value to measure the difference between the observations within a group
- Standardizing dataset, just like PCA!

K-Means Clustering



Introducing Other Methods for Future Work

- Multidimensional Scaling
  - Similar to PCA in its aim!
  - Applied to **distance matrices**
- Analysis of Repeated Measures
  - Applied to multivariate data that contains repeated measurements on the same variable on each unit
  - Aims to examine change in the repeated values of the response variable and determine any explanatory variables associated with it
  - An upgrade from usual linear regression to account for the "unobserved" effects

What I Learned/Gained!

- Many new skills!
- Project-learning experience!
- Confidence to learn on my own and craft my own projects!
- A great mentor :)