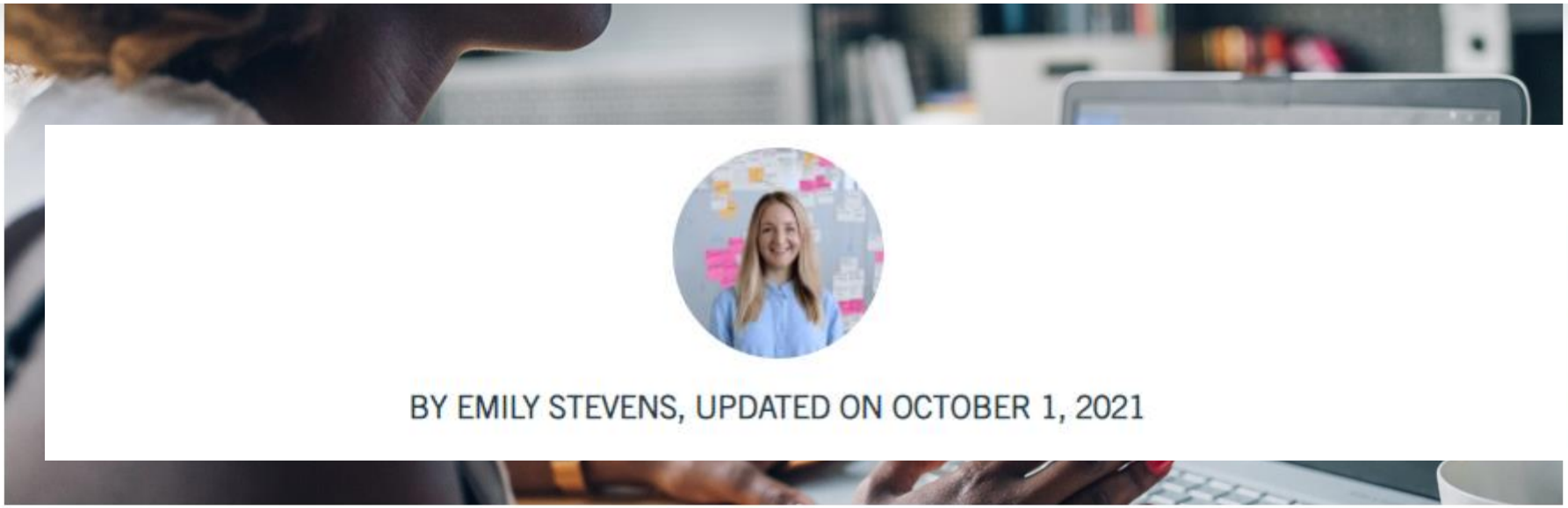
Exploratory Data Analysis: basic data set exploration



1. Confirm shape (dimension) of the dataframe
2. Confirm data types of the various columns
    - observe that our dataset has a combination of **categorical** (object) and **numeric** (float and int) features.
    - Numeric features that should be categorical and vice versa
3. Display a few rows
    - Can you understand the column names? Do they make sense? (Check with the variable definitions again if needed)
    - Do the values in these columns make sense?
    - Are there significant missing values (NaN) sighted?
    - What types of classes do the categorical features have?

4. Look at the data distribution:

What to look for:

- How the values in a feature are distributed, or how often they occur.
- For numeric features, we'll see how many times groups of numbers appear in a particular column, and for categorical features, the classes for each column and their frequency. Use both graphs and actual summary statistics. The graphs enable us to get an overall idea of the distributions while the statistics give us factual numbers. These two strategies are both recommended as they complement each other.

Numeric features

5. Plot each numeric feature

What to look for:

- Possible outliers that cannot be explained or might be measurement errors
- Numeric features that should be categorical. For example, Gender represented by 1 and 0.
- Boundaries that do not make sense such as percentage values> 100.

6. Summary statistics of the numerical features

Use functions such as,

- summary()
- describe ()

Categorical features

7. Summary statistics of the categorical features

- it is important to show the summary statistics before we plot graphs because some features have a lot of unique classes (like we will see for the Address) and the classes would be unreadable if visualized on a countplot.
- get the *count* of the values of each feature, the number of *unique* classes, the *topmost* frequent class, and how *frequently* that class occurs in the data set.
    - Countplot: histogram for categorical variables

What to look out for:

- Sparse classes which have the potential to affect a model's performance.
- Mistakes in labeling of the classes, for example 2 exact classes with minor spelling differences.

Grouping and Segmentation

Segmentation allows us to cut the data and observe the relationship between categorical and numeric features.

8. Segment the target variable by categorical features.

- Compare the target feature between the various classes of the main categorical features and see how the target changes with the classes.
- Boxplot

What to look out for: which classes most affect the target variables.

9. Group numeric features by each categorical feature.

- How other numeric features change with each categorical feature by summarizing the numeric features across the classes.
- groupby

Relationships between numeric features and other numeric features

10. Correlation matrix for the different numerical features.

- A correlation is a value between -1 and 1 that amounts to how closely values of two separate features move simultaneously. A *positive* correlation means that as one feature increases the other one also increases, while a *negative* correlation means one feature increases as the other decreases. Correlations close to 0 indicate a *weak* relationship while closer to -1 or 1 signifies a *strong* relationship.

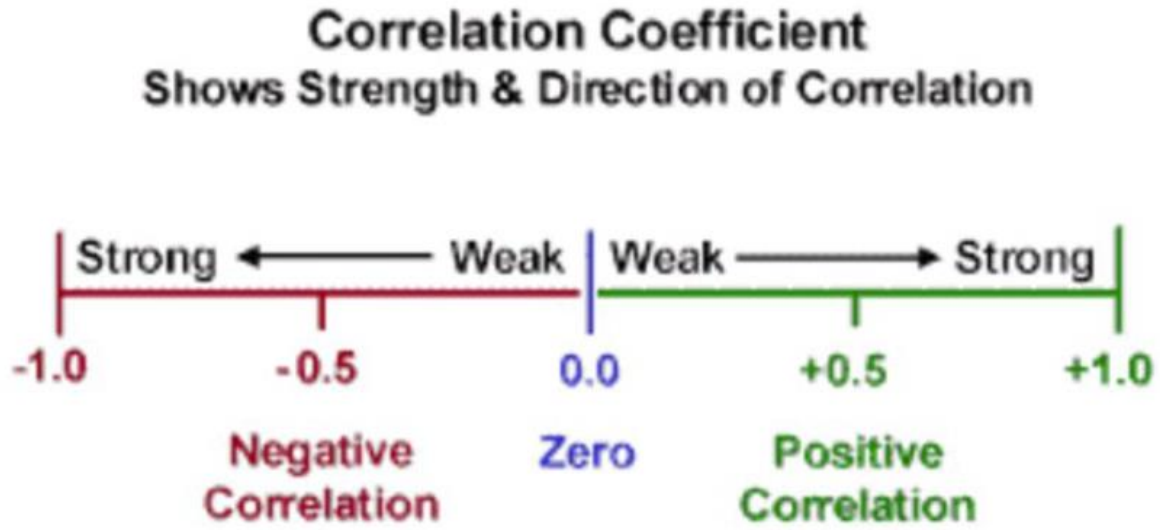


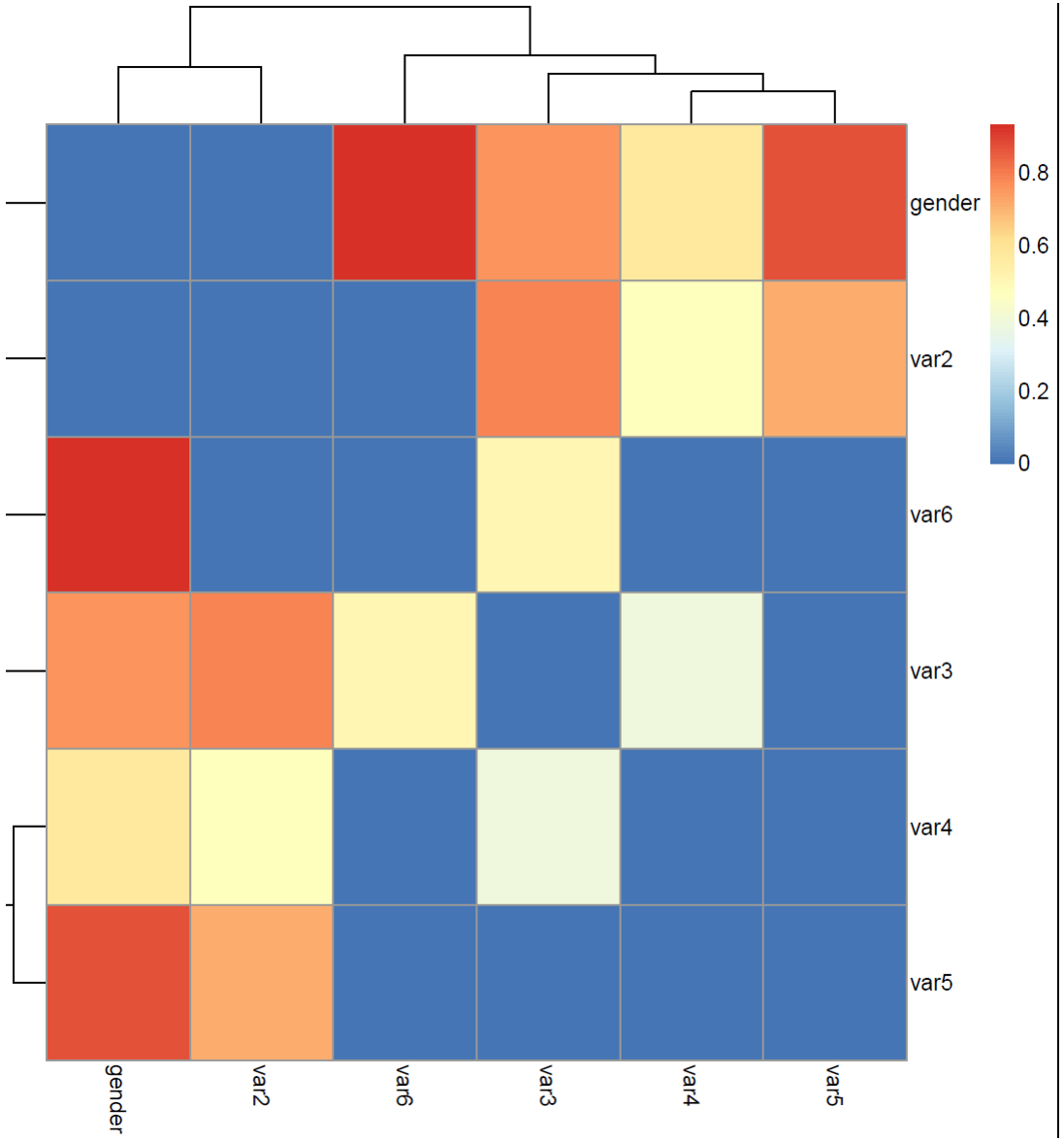
Image from edugyan.in

11. Heatmap of the correlations

A high positive correlation appears as *dark red* and a high negative correlation as *dark blue*. Closer to white signifies a weak relationship.

What to look out for:

- Strongly correlated features; either dark red (positive) or dark blue(negative).
- Target variable; If it has strong positive or negative relationships with other features.



About ggplot2: Elegant Graphics for Data Analysis (link here)

- The three key components of every plot: data, aesthetics and geoms, Section 2.3.
- How to add additional variables to a plot with aesthetics, Section 2.4.
- How to display additional categorical variables in a plot using small multiples created by faceting, Section 2.5.
- A variety of different geoms that you can use to create different types of plots, Section 2.6.
- How to modify the axes, Section 2.7.
- Things you can do with a plot object other than display it, like save it to disk, Section 2.8.