Exploratory Data Analysis (EDA)

> "Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis, and to check assumptions with the help of summary statistics and graphical representations." - Prasad Patil

Main Stages in Exploration Phase

1. Hypothesize: Form and document your initial hypotheses about what's driving the target variable. -
    - Hypothesize which features can help us determine the target variable.
    - We want to interrogate how variables interact with eachother
    - We need to explore how independent variables interact with the target variable, AKA the "y value" or the dependent variable.
2. Visualize: use visualization techniques (scatterplots, jointplot, pairgrid, heatmap) to identify drivers. When a visualization needs to be followed up with a test, we will do so.

3. Test your Hypotheses: We will analyze the drivers of a continuous variable using appropriate statistical tests (t-tests, correlation, and chi-squared hypothesis tests)

Standing Orders for the Exploration Stage

- Document your initial hypotheses. Write them down so they're concrete and not in your head.
- Document any surprises you may find in visualizing.
- Document your hypothesis test results. That means writing up when the tests reject the null hypothesis or fail to reject your null hypothesis, for each hypothesis you make.
- Write down your takeaways. Documenting your takeaways is a huge component of your final deliverable/analysis.
- Identitfy features that correlate with eachother. If feature A and feature B are each tightly correlated with the target variable, but they're also tightly correlated with eachother, we should use one feature that correlates better, rather than use both.

General Steps for Visualizations in your Explore Stage

- Always plot out the distributions. This is critical b/c many of our statisitical tools and machine learning algorithms assume certain distributions. If your data isn't remotely normally distributed, then avoid using any tools that assume normally distributed data.
- Plot out how subgroups compare to each-other and to the overall population.

Visualizing Continous to Continous

- Scatter-Matrix with seaborn `pairplot`

  ```
  import seaborn as sns
  df = sns.load_dataset("iris")
  sns.pairplot(df, hue="species")
  ```

- Scatterplot: https://seaborn.pydata.org/generated/seaborn.scatterplot.html
- Heatmap: `sns.heatmap(df.corr())`
- https://seaborn.pydata.org/generated/seaborn.lmplot.html shows a line of best fit

Visualizing Discrete x Continuous

- Swarmplot, Violinplot, Box plots
  - https://seaborn.pydata.org/examples/horizontal_boxplot.html
  - https://matplotlib.org/3.2.1/gallery/statistics/boxplot_color.html#sphx-glr-gallery-statistics-boxplot-color-py
- Bar Plots: https://seaborn.pydata.org/examples/grouped_barplot.html
- Boxplot vs. Violin example https://matplotlib.org/3.2.1/gallery/statistics/boxplot_vs_violin.html

Discrete x Discrete

- https://adataanalyst.com/data-analysis-resources/visualise-categorical-variables-in-python/

Some additoinal

- Identify if there are logical/domain/cultural cutoffs in continuous variables that would allow us to treat them as categorial values. For example, 98.45 and 99.1 are both an A or a A+ grade in most scales.
- If there's a logical cutoff point, like a grade of 70 or a voting age of 18, we can make a boolean to go along with a continuous value. This can allow us to gain additional insight in visualizing distributions between groups.
- Bin continuous variables to make categorical variables

Further EDA Resources

- https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15
- https://www.itl.nist.gov/div898/handbook/index.htm