Working on this data analytics project taught me useful skills in using R to conduct data, which can help my future career. The main technical skills I applied in this project (not including the analysis conducted by my teammates) were checking the distribution shape, testing hypotheses, and finding correlations. Using these hands-on has improved how I analyze and look at a completely new dataset in the first place and conduct exploratory data analysis (EDA).

A key lesson was seeing why I first need to check the data's distribution before picking statistics methods. Jumping right into the modeling process in a data science project without proper preparation and exploration can lead to several issues. For example, skipping the data exploration phase means you may miss important patterns, outliers, or anomalies (highly skewed data or a huge number of outlier) in the dataset.The histograms showed uneven right-skews rather than normal bell-curves. So I knew non-normal tests fit better than ones assuming symmetry. Calculating p-values also grew my understanding of what makes results significant and how to evaluate the correctness of the null hypothesis. Additionally, spotting patterns between variables with scatterplots and coefficients let me discover hidden connections such as positive correlation.

Following this process of exploring, picturing, modeling, and interpreting to find stories in the data was very useful. I learned the importance of carefully studying statistical behavior instead of jumping to conclusions. Letting the numbers guide decisions through proper analysis is valuable for data-driven choices.

In particular, skills such as using R programming language and a combination of different R packages (e.g. ggplot) help my goal of becoming a data science analyst. Identifying useful trends and insights to suggest changes requires creative, patient data work like my project. Whether examining sales, finances, web traffic, or other performance measures, thoroughly investigating the distributions and relationships is key for reliable recommendations without misusing statistics. In today’s tech community, besides Python, R is also a widely used programming language to conduct machine learning analysis. According to the Kaggle 2021 Machine Learning and Data Science Survey, R is the 2nd most popular language for data analysis/data science work, with 59% of respondents using it. Python leads at 78% usage, but R maintains a strong 2nd place position. [1]

In summary, this hands-on practice gave me reusable data abilities and critical thinking for numbers that I can apply in future roles. I now better grasp analysis as an iterative process of making and checking assumptions based on evidence. As I work with different datasets and teams, this insight will continue guiding my future career in tech. The project taught me analysis comes from grounding explanations in the numbers' signals. I look forward to sharpening these learnings of distribution-checks for methods, thorough correlations and hypothesis tests for insights, and data-first reasoning for decisions.

Reference

[1] 2021 kaggle machine learning & data science survey. (n.d.). Retrieved December 4, 2023, from Kaggle.com website:

https://www.kaggle.com/c/kaggle-survey-2021/