Capstone Project: Milestone Report

League of Legends: Ranked Games

Problem

The problem I want to solve is being able to optimize one's time spent on playing ranked matches in League of Legends to rise up in the rank ladders, to increase one's chance of winning the game simply based on the champions one chooses/bans and predict the likelihood of winning the game based on variables such as champions picked/banned on both teams, objectives and micro strategies to focus on, etc.

Client

My client is a consumer of League of Legends who is trying to go up the ranks as efficiently as possible, whether it's banning specific champions, choosing specific champions and assessing whether the combination of champions on one's team/opposing team will most likely result in a win, focusing on what areas in the game to put most energy towards, etc..

Data Wrangling Scope

I am using a dataset on Kaggle that includes 7 csv files as well as a couple of JSON files that will need to be imported, combined, manipulated, and prepared to better understand the dataset.

Rough Outline

Start off by looking at frequencies of champion picks/bans within each region, percentage win/loss ratio of each champion/side, differences in picks between regions, other variables that may affect win/loss, check for duplicate variables, check correlation between variables and target variable, look at all three Pearson's, Spearman's and Hoeffding's correlation coefficients. Use only significant variables. Look at other interesting statistics within the datasets that may draw attention.

Deliverables

Jupyter Notebook, maybe a blog post, and/or Tableau visualization.

Data Wrangling

I got the data set via REST API of League of Legends as well as several csv files that were downloaded from Kaggle. The steps for data wrangling were straightforward yet I needed to utilize several techniques to prepare the data into an interpretable format to gather insight from. Here are the steps to wrangle all the files:

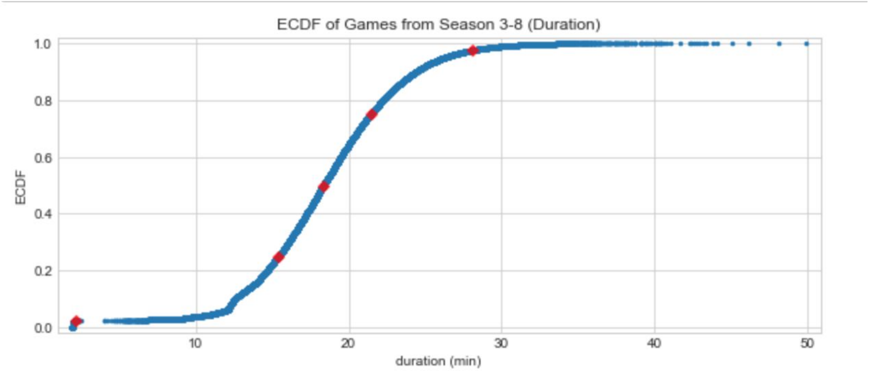
1. Extracting data from JSON file
2. Joining dataframes
3. Unioning dataframes
4. Renaming data within certain columns for clarity
5. Converting Data within certain columns
6. Changing Column Types
7. Reindexing

8. Selecting only useful columns for conciseness

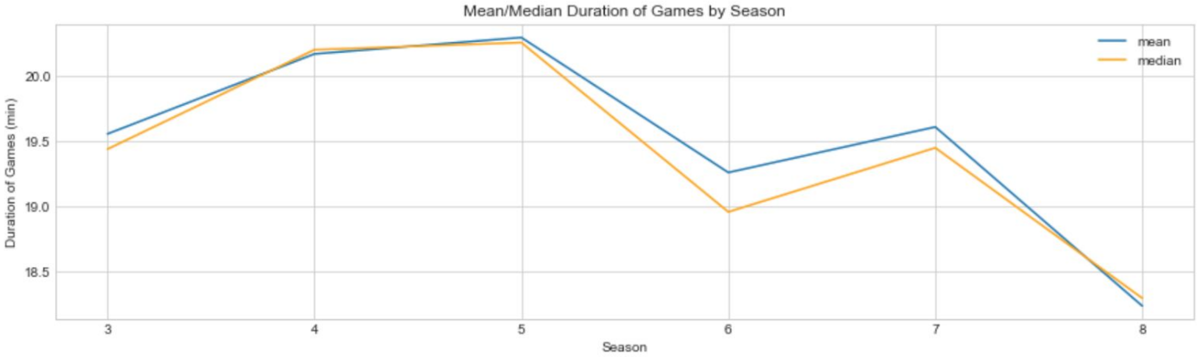
9. Renaming columns for clarity

Exploratory Data Analysis

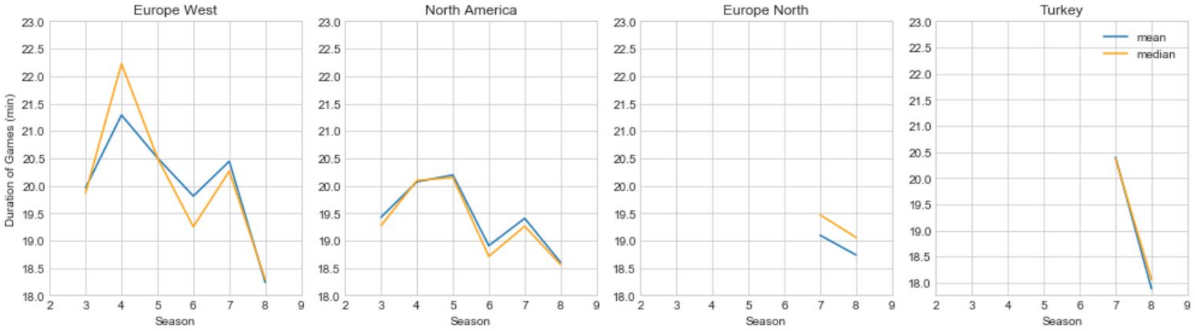
Looked at duration of all season through an ECDF plot and identified through further exploration that duration follows a relatively normal distribution.



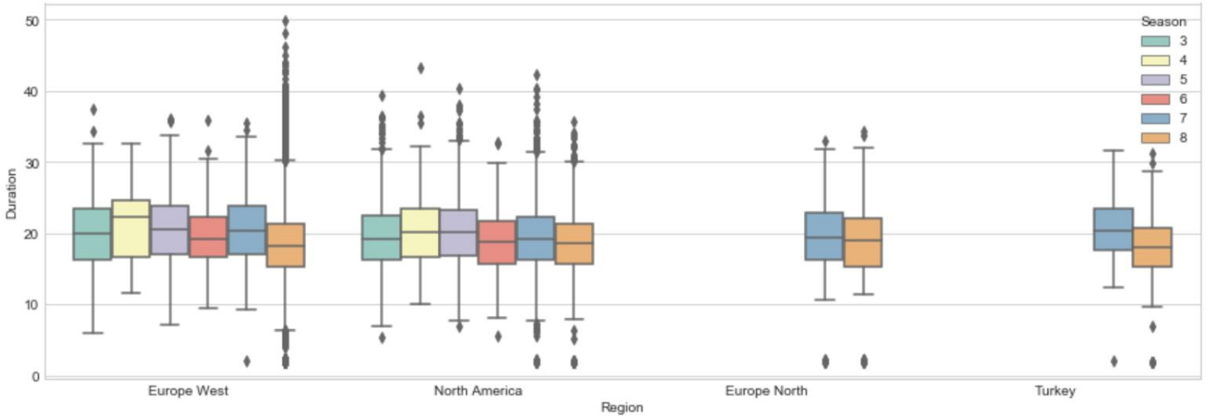
Looked at the mean and median durations by season and found a bit of a difference in season 6 and 7 which can be points to delve into.



Looked at the different regions and its respective means/median durations and we see the most radical behavior in Europe and West which is quite interesting. Europe North and Turkey are not quite mature yet but in the future, I would use North America as a standard sample whereas, I would use Europe West to dig into even further.

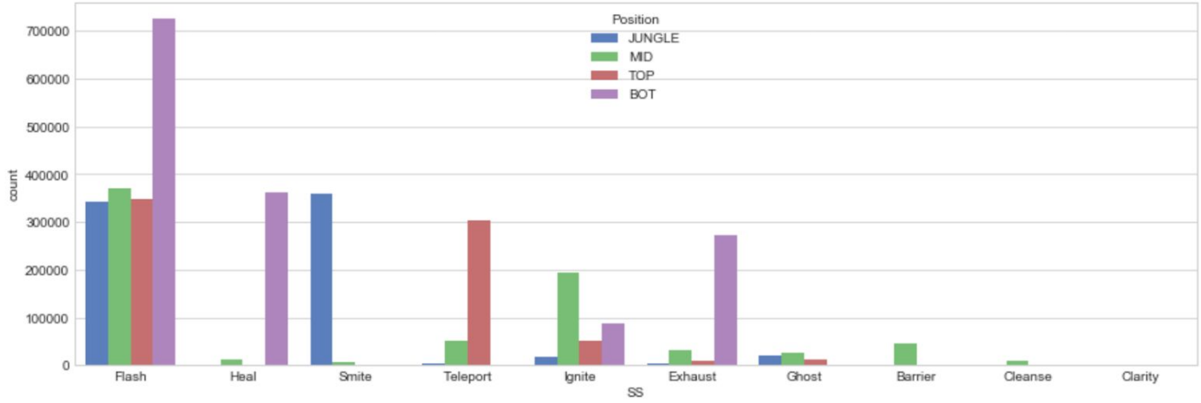


By digging a little further, we further strengthen our speculation of Europe West being unstable unstable every season as well as a significant number of outliers which makes the dataset even more worth to look into.

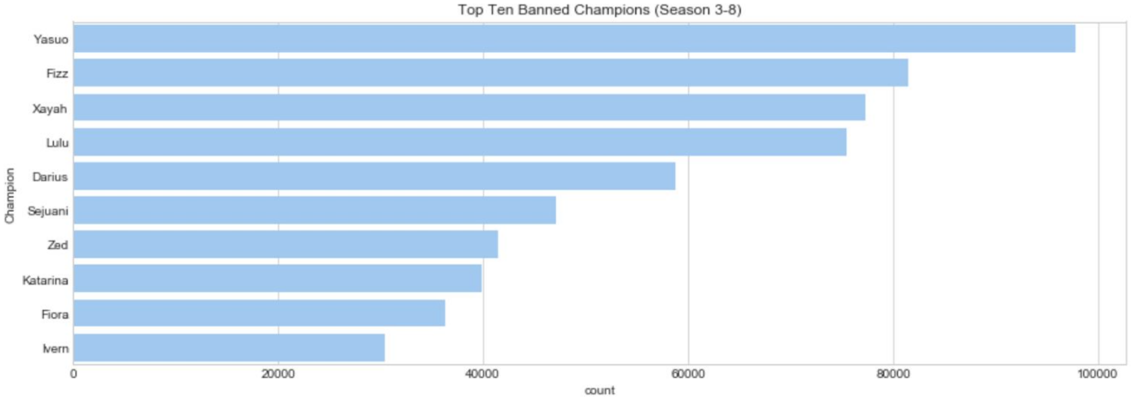


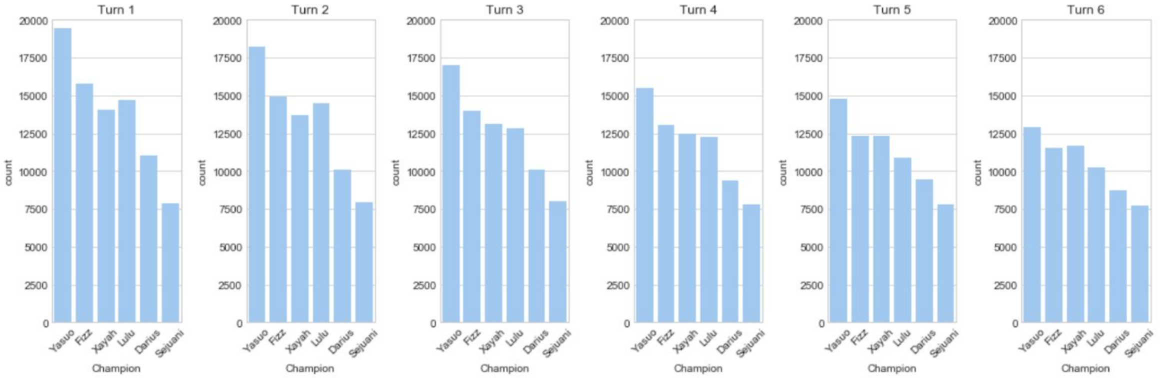
By looking at the durations in different perspective, we may hypothesize and speculate different strategies, metas, and gameplay to take into consideration when giving statistically sound suggestions to players. Each region will have different features that players will need to cater their skillset and gameplay towards as a result. Essentially, significance in one region may tend to not be as effective in another.

Here is a simple chart that shows the most prominent summoner spells that each position uses. This chart may also be used to predict percentage-wise which summoners spells your opponent role will most likely pick and adjust your own summoner spells accordingly.



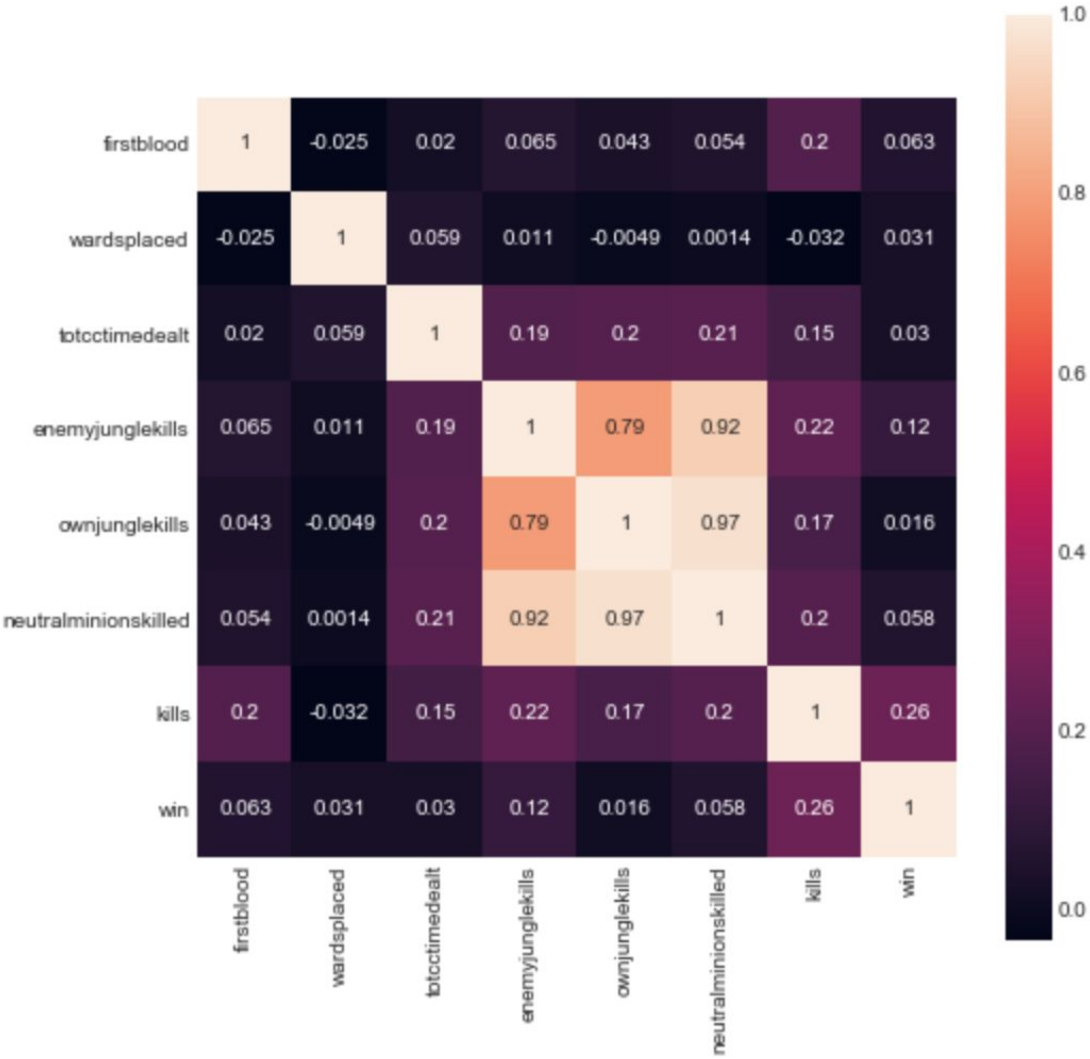
Below, it’s quite amazing seeing the top 6 banned champions from season 3-8 are in the same exact order in terms of ban turn order. This shows great significance in the the top ten banned champs. This can then be leveraged when choosing which champions to master when training to play in ranked games. It might not be extremely optimal to spend hundreds of hours practicing a champion that’s heavily banned in ranked games.





As you can see below, no variables in the micro game have a high correlation with a win in League of Legends. The only high correlation within the heat map is between dependent variables and thus, should be addressed as duplicate variables. When you kill a neutral

minion, you are also technically killing an enemy jungle minion as well as your own jungle minion. Thus, the high correlation in the heat map should be ignored.



Inferential Statistics

Each game in League of Legends has a *blue* and *red* side. Although each side is **randomly assigned** for each game, each side has it's **differences** which may or may not be leading to advantages before the game even starts.

Blue side has it's red buff on the **bottom side** (near their duo lane) whereas red side has it's red buff on the **top side** (near their top lane). Certain junglers who heavily rely on their mana in the jungle tend to start their blue buff **first** and getting randomly assigned to a red/blue side will dictate where the jungler **finishes** their early game jungle route (there will always be variations depending on jungler/strategy).

This is noteworthy because not only will **respective lanes be affected**, but also objectives (towers, dragons, scuttle crab, rift herald, etc.) will be closer in proximity / on route for certain sides and junglers.

With lanes and objectives impacted solely on randomly assigned sides for teams, I am curious and even doubt if the proportion of wins on the blue side is equal to the proportion of wins on the red side.

After performing a two sample z-test on the difference between our sample proportions, our statistical analysis says that we are *confident* that the true difference between winning percentages on blue and red side is between 0.005 and 0.021, 99% of the time. Thus, we are quite confident that the true difference is between 0.005 and 0.021 **BUT**that percentage is **AT MOST** less than 2%.

Furthermore, our analysis **ONLY** looked at the relationship between assigned sides and winning percentages. There are a couple more variables to look and to uncover the correlation between such as win percentages of certain champions and even looking at success rate of all the combinations of champions (which may get quite complex).

Thus, if I were making suggestions to a player seeking to rise up the ranking ladder, I would not advise him to leave a game if randomly sorted to a specific side based on statistical **AND** practical significance.

Concluding Milestone Report

Through EDA and performing inferential statistics alone, suggestions can be made to our client that can significantly help optimize their climb up the ranking ladders in terms of time, strategy, gameplay and practice mentality. There are also quite interesting facts and findings, with further investigation, that can possibly give players an edge over opposing players. With only several techniques applied in preparing, exploring and testing our data set, I can’t wait to uncover deeper insight to help players prepare for the climb to the top with data science!