Which variables of a match of Dota 2 most correlate a team's GPM?
Project Members: Saidel Halol, Arthur Tran
COGS 109, Fall 2019

I. Introduction.

Dota 2 is a multiplayer online battle arena (MOBA) video game developed and published by Valve Software. Released in 2013 as a standalone version of the popular Warcraft 3 custom game, Dota 2 has quickly become one of the most popular video games in the world due to its complexity and competitiveness. The game pits two teams of five players against each other with the main goal to destroy the other team's "ancient": a large structure deep in a team's base. Each player controls a unique "hero" character, which has powerful abilities and different strengths and weaknesses. Connecting each of the teams bases are three lanes where weaker, A.I. controlled units called "creeps" travel down each lane from each team's base, which spawn from buildings called "barracks" inside the team's base. If a team's barracks are destroyed, the opposing team's creeps will get stronger. Guarding the lanes for both teams, are stationary structures called "towers" which provide vision for their team and fire upon any opposing creep or hero units. Heroes can kill these creeps, towers, and enemy heroes to gain gold and experience points (exp). Gold is used to buy powerful items which heroes can use to enhance their stats like health, mana, or damage, provide powerful abilities to cripple their opponents, or protect their allies and themselves from the powerful abilities of the enemy. Exp is used to increase a hero's level, and when a hero levels up, their stats and abilities become stronger. If a hero's health reaches 0, they die and are taken out of the game for a certain amount of time, which is longer the higher the level of the dead hero's is. When a team feels they have an advantage, that team will want to fight the other team, and the winner of the resulting engagement, called a "teamfight" will want to push their advantage and kill creep waves and towers in a lane or multiple lanes in order to get closer to the opponent's ancient.

There are many variables that go into why a team won a match of Dota, and not the other team. Things like a team having a higher gpm (gold per minute) and xpm (experience points per minute) as well as team total kills and assists. GPM is typically the variable that tells which team is winning in a match of Dota. People casting a game of Dota will typically refer to gpm and gold difference to get a sense on if a team is winning (though this may not always be the case). If a team has a higher gold difference and gpm than the other they will have an advantage, because that team will have access to more powerful items and are more favored in teamfights [1]. In this project, we aim to find which aspects of the game most correlate a team's gpm. The dataset *Dota 2 Matches* on kaggle.com by Devin Anzelmo is a collection of 50,000 matches of Dota 2 from 2016, recording important features of each match like items bought, heroes picked, total gold, the game's all chat, etc [2]. Using this dataset, we are able to see the gpm differences between winning teams and losing teams and see if variables like assists, deaths, and last hits correlate with a higher gpm. In conducting this project, we hypothesize that team based statistics like gpm and xpm most correlate with a higher chance of a team winning and that winning teams will typically have more last hits and assists than the losing team but have less deaths.

II. Methods

Using Python via Jupyter Notebook, we analyzed the impact of certain statistics' impact on the outcome of a match. Since team gold allows the purchase of items to obtain relative strength through boosts in statistics, we determined that gpm is the main statistic that dictates the pace of the match since the more gold a team has, the more items each hero can buy and the stronger the team becomes. We split up our data into two dataframes, with one containing the statistics of the winning teams and the other, the statistics of the losing teams in order to find and plot trends that might exist for winning and losing teams respectively. In our regression models, we have gpm be the factor that our predictors such as number of last hits, team assists, and deaths, are impacting. We used logistic and polynomial regressions to predict how much a team's gold per minute accumulation is impacted and therefore predict the team's chances of winning the match.

III. Results

A. Model Selection

For our models, we decided to use logistic regressions to determine whether gpm can be used as a predictor for match wins. After graphing a scatter plot that depicts average team gpm of winning and losing teams we can see that winning teams typically have a higher average team gpm by the end of the game, showing that gpm is a telling factor of whether or not a team will win a match (Figure 1). When we plotted the regression for assists as a function of gpm, there is a slightly negative correlation for the winning teams, but the total gpm is higher while the trend for losing teams is that there is a positive correlation, but the total gpm is lower (Figure 2). A possible explanation for this phenomenon is that there is a comeback mechanic that awards more gold for kills when a team is very behind meaning more gold is distributed for assists and explains why more assists on the losing team contributes more to the team's gpm but the teams' total gpm is still low. When we plotted the regressions for last his as a function of gpm, there is a slight negative correlation for the winning teams, but there is a positive correlation for the losing teams (Figure 2). This could be because the winning teams typically have higher kill counts than the losing teams meaning that creep farming could be less impactful to their gpm than teams who are losing where farming might be their primary source of income during the game. Plotting the regressions for deaths as a function of gpm, we found that there is a negative correlation for the winning teams showing that dying many times netted a loss in gpm, however, there is a positive correlation for the losing teams (Figure 2). This can be attributed to match length, where although the losing team dies a lot, as the game progresses, the losing team will still become strong enough to make up for lost gold through other means. Finally we performed polynomial regressions using OLS for all three variables as functions of gpm, and found that in winning teams, the statistically significant factors were last hits and deaths, while in losing teams all three variables were statistically significant (Figure 3).

B. Model Estimation

For the final parameter estimates of our linear regression models,we obtained -0.3627 as the coefficient for assists in winning teams, and 1.4013 for the losing teams (Figures 4 and 5). We obtained -0.0292 as the coefficient for last hits in winning teams and 0.1256 for losing teams. We obtained -1.1107 as the coefficient for deaths in winning teams and 1.8391 for losing teams. For our two polynomial models, the coefficients of the winning teams' variables were, -0.0047 for assists, 0.0116 for last hits, and -1.2307 for deaths. The coefficients for the losing teams' variables were 0.7709 for assists, 0.0799 for last hits, and 0.3307 for deaths. All of these parameters show small coefficients which might suggest that on their own each parameter doesn't impact teams' gpm very much. To find the accuracy of each model's predictions we can look at the R-squared values of each OLS model. The R-squared values for assists as a function of gpm is 0.033 for the winning teams and 0.457 for the losing teams. The R-squared values for last hits as a function of gpm is 0.026 for winning teams and 0.474 for the losing teams. The R-squared values for deaths as a function of gpm is 0.115 for the winning teams and 0.165 for the losing teams. For the polynomial OLS models that use all three parameters as predictors, the R-squared value is 0.118 for the winning teams and 0.583 for the losing teams. As we can see the R-squared values of many of the regression models are low, meaning that the models aren't very accurate for many of the tests. However an explanation for this can be that there are many variables that are accounted into a match that no one statistic on its own can accurately predict the outcome of the game.

IV. Conclusions and Discussion

Based on the results from our analyses, we can conclude that although assists, last hits, or deaths contribute to determining a team's average gpm, which has the ability to predict a team's chances of winning, no one variable is a solid predictor of whether or not a team will win. Some potential next steps people working with this dataset can do is take into account other variables like duration of the match and the heroes picked in the game. Having a larger duration of a match can mean larger values for variables like team assists, team deaths, and gpm (since the gold bounty of creeps increases as time goes on). Hero synergy in the picking/drafting hero phase before a match is very important because picking an unbalanced team (i.e. more damage heroes and no support heroes) will typically be at a disadvantage than a balanced team, though this may be difficult to measure as some heroes can be played multiple ways (damage/core and support) and balance patches may bring into light newer ways to play a hero or make a hero's playstyle unfavorable and "out of the meta."

V. References

[1] Dota. "Gold." *Dota 2 Wiki*, Gamepedia, 30 Nov. 2019, https://dota2.gamepedia.com/Gold.

[2] Anzelmo, Devin. "Dota 2 Matches." *Kaggle*, 2016, https://www.kaggle.com/devinanzelmo/dota-2-matches.

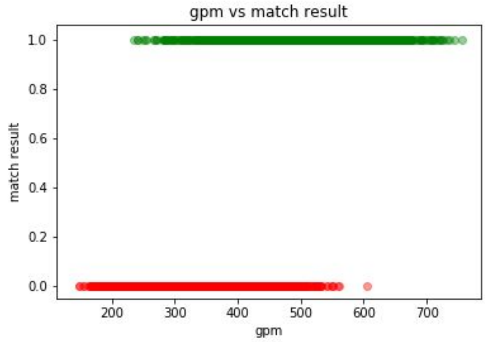

Figure 1: This is a graph showing team average gold per minute (GPM) for winning teams (green) vs losing teams (red).

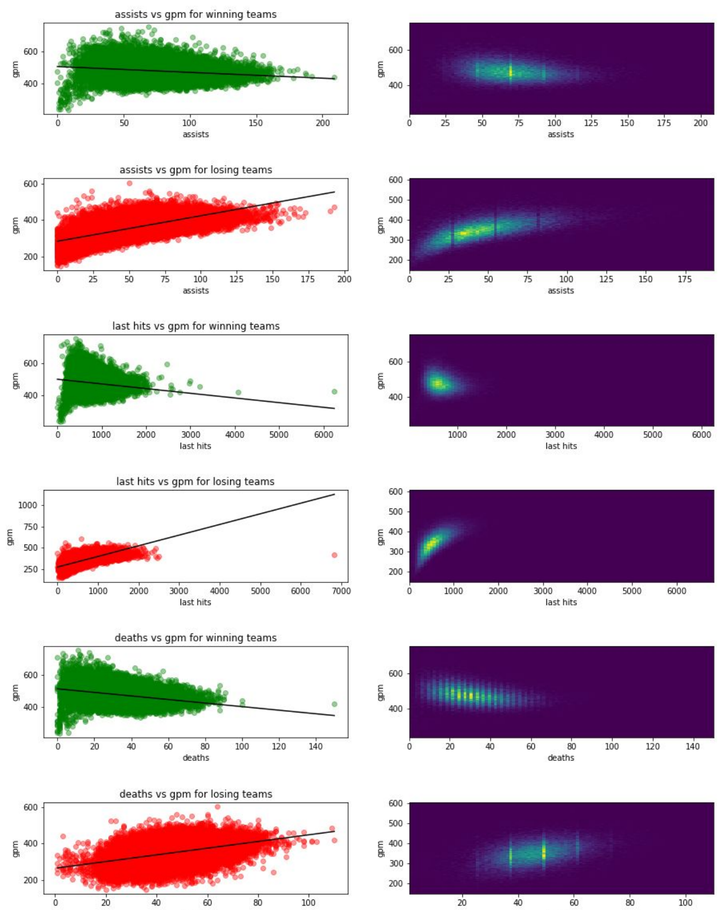

Figure 2: Series of graphs and 2d histograms depicting team assists, last hits, and deaths for winning and losing teams vs GPM.

OLS Regression Results

Dep. Variable:	gpm	**R-squared:**	0.118
Model:	OLS	**Adj. R-squared:**	0.118
Method:	Least Squares	**F-statistic:**	2231.
Date:	Thu, 05 Dec 2019	**Prob (F-statistic):**	0.00
Time:	14:52:01	**Log-Likelihood:**	-2.5759e+05
No. Observations:	50000	**AIC:**	5.152e+05
Df Residuals:	49996	**BIC:**	5.152e+05
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	511.2212	0.724	706.351	0.000	509.803	512.640
assist	-0.0047	0.010	-0.479	0.632	-0.024	0.015
lasthits	0.0116	0.001	12.235	0.000	0.010	0.013
deaths	-1.2307	0.019	-63.968	0.000	-1.268	-1.193

Omnibus:	1693.669	**Durbin-Watson:**	1.997
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	3883.060
Skew:	0.193	**Prob(JB):**	0.00
Kurtosis:	4.309	**Cond. No.**	2.85e+03

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.85e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

OLS Regression Results

Dep. Variable:	gpm	**R-squared:**	0.583
Model:	OLS	**Adj. R-squared:**	0.583
Method:	Least Squares	**F-statistic:**	2.326e+04
Date:	Thu, 05 Dec 2019	**Prob (F-statistic):**	0.00
Time:	12:46:15	**Log-Likelihood:**	-2.4420e+05
No. Observations:	50000	**AIC:**	4.884e+05
Df Residuals:	49996	**BIC:**	4.884e+05
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	249.3218	0.643	388.003	0.000	248.062	250.581
assist	0.7709	0.008	93.052	0.000	0.755	0.787
lasthits	0.0799	0.001	120.211	0.000	0.079	0.081
deaths	0.3307	0.015	21.750	0.000	0.301	0.360

Omnibus:	4289.075	**Durbin-Watson:**	2.007
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	18587.055
Skew:	0.330	**Prob(JB):**	0.00
Kurtosis:	5.913	**Cond. No.**	2.99e+03

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.99e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

Figure 3: Polynomial regression results for winning teams (left) and losing teams (right) for the three variables: assists, last hits, and deaths.

OLS Regression Results

Dep. Variable:	gpm	**R-squared:**	0.033
Model:	OLS	**Adj. R-squared:**	0.033
Method:	Least Squares	**F-statistic:**	1728.
Date:	Thu, 05 Dec 2019	**Prob (F-statistic):**	0.00
Time:	12:46:17	**Log-Likelihood:**	-2.5988e+05
No. Observations:	50000	**AIC:**	5.198e+05
Df Residuals:	49998	**BIC:**	5.198e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	506.5736	0.645	785.413	0.000	505.309	507.838
assist	-0.3627	0.009	-41.574	0.000	-0.380	-0.346

Omnibus:	1730.755	**Durbin-Watson:**	1.996
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	3341.679
Skew:	0.263	**Prob(JB):**	0.00
Kurtosis:	4.152	**Cond. No.**	244.

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

OLS Regression Results

Dep. Variable:	gpm	**R-squared:**	0.026
Model:	OLS	**Adj. R-squared:**	0.026
Method:	Least Squares	**F-statistic:**	1322.
Date:	Thu, 05 Dec 2019	**Prob (F-statistic):**	8.69e-286
Time:	12:46:18	**Log-Likelihood:**	-2.6008e+05
No. Observations:	50000	**AIC:**	5.202e+05
Df Residuals:	49998	**BIC:**	5.202e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	501.1662	0.588	852.711	0.000	500.014	502.318
lasthits	-0.0292	0.001	-36.364	0.000	-0.031	-0.028

Omnibus:	1805.628	**Durbin-Watson:**	1.996
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	3170.634
Skew:	0.305	**Prob(JB):**	0.00
Kurtosis:	4.073	**Cond. No.**	2.19e+03

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.19e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

OLS Regression Results

Dep. Variable:	gpm	**R-squared:**	0.115
Model:	OLS	**Adj. R-squared:**	0.115
Method:	Least Squares	**F-statistic:**	6524.
Date:	Thu, 05 Dec 2019	**Prob (F-statistic):**	0.00
Time:	12:46:19	**Log-Likelihood:**	-2.5767e+05
No. Observations:	50000	**AIC:**	5.153e+05
Df Residuals:	49998	**BIC:**	5.154e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	515.2142	0.463	1113.132	0.000	514.307	516.121
deaths	-1.1107	0.014	-80.772	0.000	-1.138	-1.084

Omnibus:	1708.360	**Durbin-Watson:**	1.997
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	4051.059
Skew:	0.181	**Prob(JB):**	0.00
Kurtosis:	4.347	**Cond. No.**	83.3

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Figure 4: Linear regression results for winning teams for the three variables: assists (left), last hits (center), and deaths (right).

OLS Regression Results

Dep. Variable:	gpm	**R-squared:**	0.457
Model:	OLS	**Adj. R-squared:**	0.457
Method:	Least Squares	**F-statistic:**	4.202e+04
Date:	Thu, 05 Dec 2019	**Prob (F-statistic):**	0.00
Time:	12:46:20	**Log-Likelihood:**	-2.5079e+05
No. Observations:	50000	**AIC:**	5.016e+05
Df Residuals:	49998	**BIC:**	5.016e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	283.0574	0.361	783.136	0.000	282.349	283.766
assist	1.4013	0.007	204.995	0.000	1.388	1.415

Omnibus:	1512.534	**Durbin-Watson:**	2.012
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1962.605
Skew:	0.353	**Prob(JB):**	0.00
Kurtosis:	3.666	**Cond. No.**	117.

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

OLS Regression Results

Dep. Variable:	gpm	**R-squared:**	0.474
Model:	OLS	**Adj. R-squared:**	0.474
Method:	Least Squares	**F-statistic:**	4.512e+04
Date:	Thu, 05 Dec 2019	**Prob (F-statistic):**	0.00
Time:	12:46:21	**Log-Likelihood:**	-2.4997e+05
No. Observations:	50000	**AIC:**	4.999e+05
Df Residuals:	49998	**BIC:**	5.000e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	273.3203	0.391	698.137	0.000	272.553	274.088
lasthits	0.1256	0.001	212.412	0.000	0.124	0.127

Omnibus:	4385.619	**Durbin-Watson:**	2.003
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	30040.770
Skew:	0.043	**Prob(JB):**	0.00
Kurtosis:	6.796	**Cond. No.**	1.62e+03

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.62e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

OLS Regression Results

Dep. Variable:	gpm	**R-squared:**	0.165
Model:	OLS	**Adj. R-squared:**	0.165
Method:	Least Squares	**F-statistic:**	9854.
Date:	Thu, 05 Dec 2019	**Prob (F-statistic):**	0.00
Time:	12:46:22	**Log-Likelihood:**	-2.6155e+05
No. Observations:	50000	**AIC:**	5.231e+05
Df Residuals:	49998	**BIC:**	5.231e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	264.5284	0.876	301.849	0.000	262.811	266.246
deaths	1.8391	0.019	99.266	0.000	1.803	1.875

Omnibus:	130.383	**Durbin-Watson:**	2.012
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	167.877
Skew:	-0.014	**Prob(JB):**	3.52e-37
Kurtosis:	3.283	**Cond. No.**	205.

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Figure 5: Linear regression results for losing teams for the three variables: assists (left), last hits (center), and deaths (right).