**Justin Thomas**

**Data Boot Camp**

**HW1 Excel**

HW Task 1 Golf Game

**HW Task3 Filtered Games Sales**

**HW Task 4**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT | |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
| *Regression Statistics* | |  |  |  |  |  |  |  |  |
| Multiple R | 0.905397 |  |  |  |  |  |  |  |  |
| R Square | 0.819744 |  |  |  |  |  |  |  |  |
| Adjusted R Square | 0.819135 |  |  |  |  |  |  |  |  |
| Standard Error | 199.9379 |  |  |  |  |  |  |  |  |
| Observations | 595 |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |  |  |  |
| Regression | 2 | 1.08E+08 | 53811072 | 1346.112 | 5.535E-221 |  |  |  |  |
| Residual | 592 | 23665307 | 39975.18 |  |  |  |  |  |  |
| Total | 594 | 1.31E+08 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |  |
| Intercept | 3.560075 | 51.9345 | 0.068549 | 0.945371 | -98.43819783 | 105.5583 | -98.4382 | 105.5583 |  |
| Age | -3.24475 | 1.928635 | -1.68241 | 0.093017 | -7.032545096 | 0.543052 | -7.03255 | 0.543052 |  |
| Minutes Played | 0.510558 | 0.009848 | 51.84526 | 3E-222 | 0.491217338 | 0.529899 | 0.491217 | 0.529899 |  |
|  |  |  |  |  |  |  |  |  |  |

**Analysis of NBA Stats Regression**

I chose to use Age and Minutes played as my independent variables and theorized that neither of them were reliably correlated to points during a game. I also theorized that minutes played would be more closely correlated to points than age even if the relationship was not statistically reliable.

The ‘R Square” value obtained is ~.81 which means that ~81% of the variation around the dependent variable Points is explained by the independent variables. Looking further at the analysts the significance of F is a very high number which indicates that there is a high probability that the regression correlation is not statistically significant/was obtained by chance vs. an actual correlation. This indicates that the two independent variables are not statically reliable indicator of the number of points scored. The coefficient for age is -3.24475 which indicates as the value of age increased the points scored decreases. The coefficient for minutes played is .5105 which indicates that as the value of minutes played increases the points scored also increases although the correlations are not overall very strong with a high probability of being impacted by random chance. The graphs below show that there is a higher linear correlation between points and minutes played than points and age. This was a good data set to show a weak correlation between points and another variable but not a good example of a statically reliable relationship between either of the independent variables and the dependent variable.

**HW Part 2 Kickstart My Chart**

1. Given the provided data, what are three conclusions we can draw about Kickstarter campaigns?

* Using the pivot analysis conducted on state by category it is apparent that the Kickstarter category Music has the highest rate of success with 79.41% of campaigns being successful. The least successful category of Kickstarter is journalism with all campaigns cancelled over the observed period for a 0% success rate.
* Using the pivot analysis conducted on subcategories that the subcategory of music does impact Kickstarter campaign success with multiple categories of music having a zero percent success rate. In terms of raw numbers rock and indie rock had the most successful Kickstarter campaigns although multiple categories had a 100% success rate even if the number of campaigns lagged behind rock and indie rock.
* Using a pivot analysis of Kickstarter campaigns over time using the “Date Created Conversion” is apparent that Kickstarter peaked in popularity/usage in 2015 and begin a steep decline in usage to the end of the dataset in 2017 which is typical of many online technology initiatives.

1. **What are some limitations of this dataset?**
   1. This data set is not filtered for outliers such as campaigns where fundraising far exceeded its goals so care must be taken when conducting analysis to filter results and or adjust calculations of statistics such as averages/means to compensate for outliers.

**3. What are some other possible tables and/or graphs that we could create?**

Many other analyses can be conducted to understand the efficacy of Kickstarter campaigns and what factors influence success or failure. For example, a simple of analysis of average of percent funded vs. state below would be useful if additional manipulation was conducted on the underlying data to compensate for outliers. In addition, regression analysis could be conducted on numerical factors such as number of backers impacted success or failure rates. Non numerical factors such as country or spotlight status could also be subjected to pivot analysis to see their impact on Kickstarter success as well as overall contribution to the makeup of the Kickstarter campaigns.