# D214 – Capstone: Executive Level Summary

## Lucas Benish – 009978338

## Western Governors University

## Develop an executive summary using your data and results from task 2. The summary should be written for a *technical audience* of your data-analytics peers or members of your team and should include *each* of the following:

### Problem statement/research question and hypothesis

The research question I proposed is as follows: utilizing a given dataset of video game data, to what extent does a video game’s price affect its users’ play time?

This question was to be answered by either accepting or rejecting the following hypotheses:

H0: The price of a game does not statistically significantly (p<0.05) affect median play time.

HA: The price of a game does statistically significantly (p<0.05) affect median play time.

### Summary of data analysis process

The data analysis process followed these outlined steps:

1. Import packages necessary to perform analysis.
2. Import data from .csv file.
3. Assess structure/missingness/duplicates of dataset.
4. Convert price data into USD.
5. Restructure/transform columns as necessary.
   1. Split columns with multiple strings into multiple columns.
6. Determine presence of outliers.
7. Visualize patterns of variables.
   1. Univariate
   2. Bivariate
   3. Multicollinearity
8. Set random seed.
9. Create initial simple linear regression model, intercept only.
10. Create “full” linear regression model, all explanatory variables of interest.
11. Create “better” linear regression model utilizing stepwise approach.
12. Calculate error measures.
    1. RMSE
    2. AIC
13. Visualize predictions and actual values.
14. Share findings/insights.

### Outline of findings

The following are the high-level findings from the analysis performed:

* Price was not a statistically significant variable in the final model predicting median playtime.
  + Therefore, the null hypothesis is not rejected.
* Positive and negative reviews, as well we game release date, were the significant variables included in the model.
* RMSE was worse for the “better” model than the “full” model.
* AIC was improved for the “better” model than the “full” model.
* Predicted values demonstrate less variability than the actual values, likely due to reduction of AIC being driving factor for model creation.

### Explanation of limitations of the techniques and tools used

The following are limitations of the techniques utilized:

* Linear regression models assume that independent variables are not correlated to each other, there was a moderate correlation between two of my independent variables.
* Data was only utilized when the release year was > 2018, so the model may have been different if different inputs were utilized.

### Summary of proposed actions

Proposed actions:

* Continue to gather data, perform model creation with newer games to see how current model compares.
* Create buckets/bins for categorical variables to allow for inclusion in full model.
  + Developer
  + Publisher
  + Steamspy\_tags

### Expected benefits of the study

Benefits of this study:

* Improved understanding of what factors may affect a users’ playtime of a specific game.
* Consumers able to predict their expected playtime based on significant variables.