KOBE’s Shot percentages

# Introduction

Many types of statistical modeling are used for different applications. For normally distributed data pulled from a single, or related variables there is the T-tests, linear regression, permutation test and more. When multiple groups are present we have ANOVA and MANOVA tests. There are ways to cut down on the number of variables to get an efficient model that does not “over fit” our data. But the limitations of these methods is the response variables has to be numeric, or analog. To describe a categorical variable, or digital, there is logistical regression. We will be running logistical regression on Kobe Bryant’s shots trying to predict if he will make or miss the shot. We will also be answering some questions about different scenarios and how they affect his shot percentage.

# Data Description

Using data provided by Kaggle for free at [**https://www.kaggle.com/c/kobe-bryant-shot-selection**](https://www.kaggle.com/c/kobe-bryant-shot-selection). The data consist of 29 years of Kobe Bryant’s shot history: 25 variables and roughly 30 thousand observations. The data set includes 5 thousand observations with an unknown value for the shot\_made\_flag response variable. To prevent leakage, predictive models must use only the observations that came before the missing value to predict the missing value.

# Exploratory Data Analysis (EDA)

Before starting on a project, we like to know the data and see what is useful and what we need to transform, because in most statistical models there are many assumptions that need to be meet. Logistical regression does not have these same limitations, the main assumptions focus on independence of variables, which we can assume, because Kobe cannot shoot the same shot twice. The purpose of the EDA is to visualize the different variables and their relationship to Kobe’s shot percentage. We used box plots to look at the different numeric variables and frequency counts on the different categorical variables. In fig.1 vs fig.2 we are displaying two different distributions.

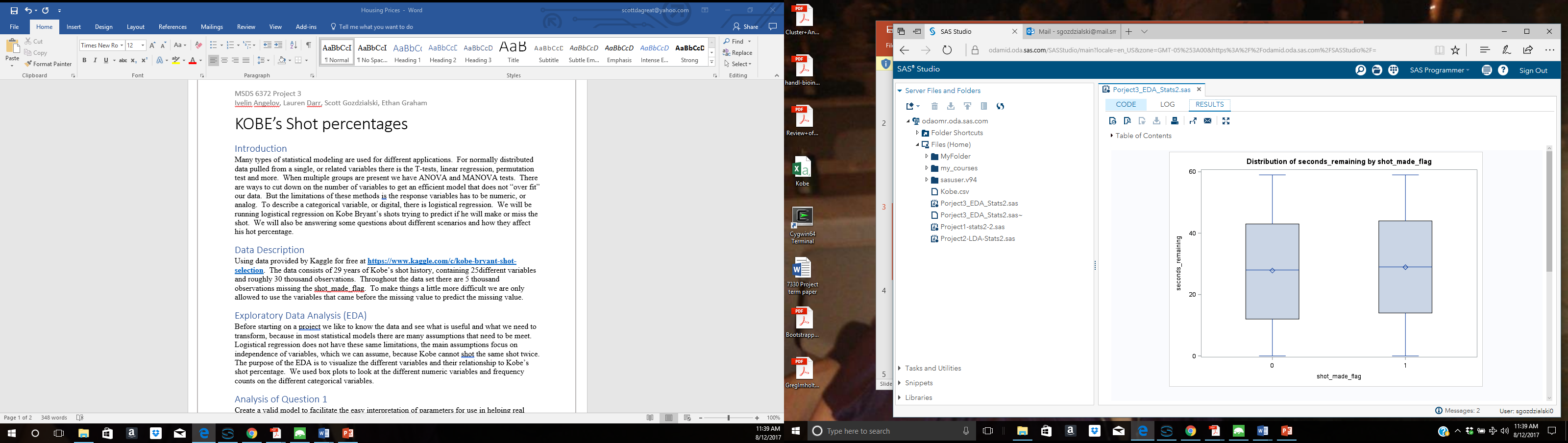


Fig. 1 (missed, left made, right)

Seconds remaining

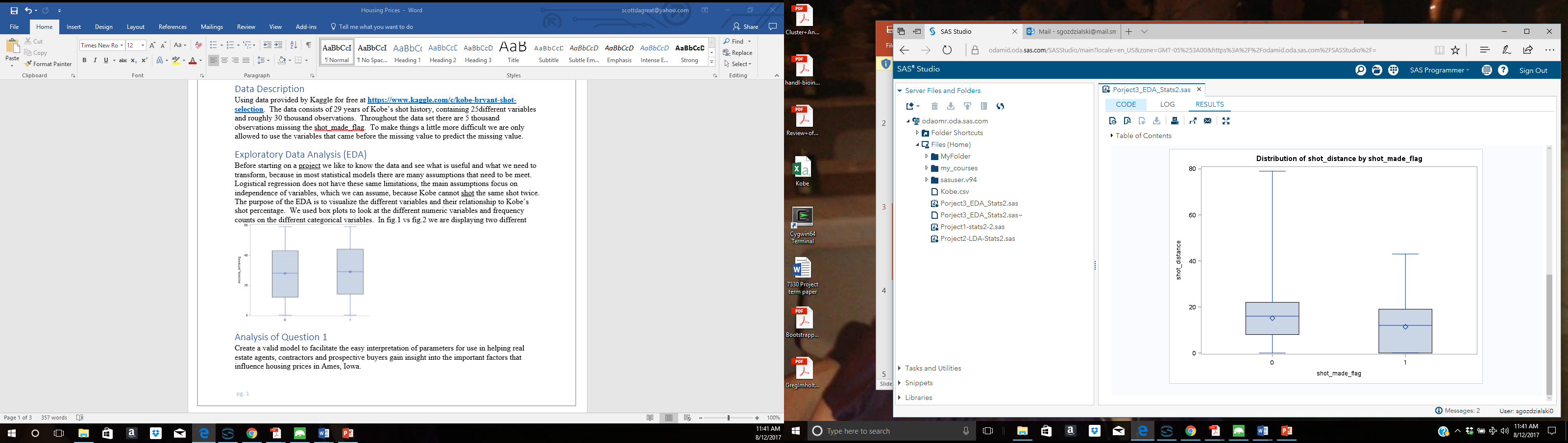


Fig. 2 (misses, left made, right)

shot distance

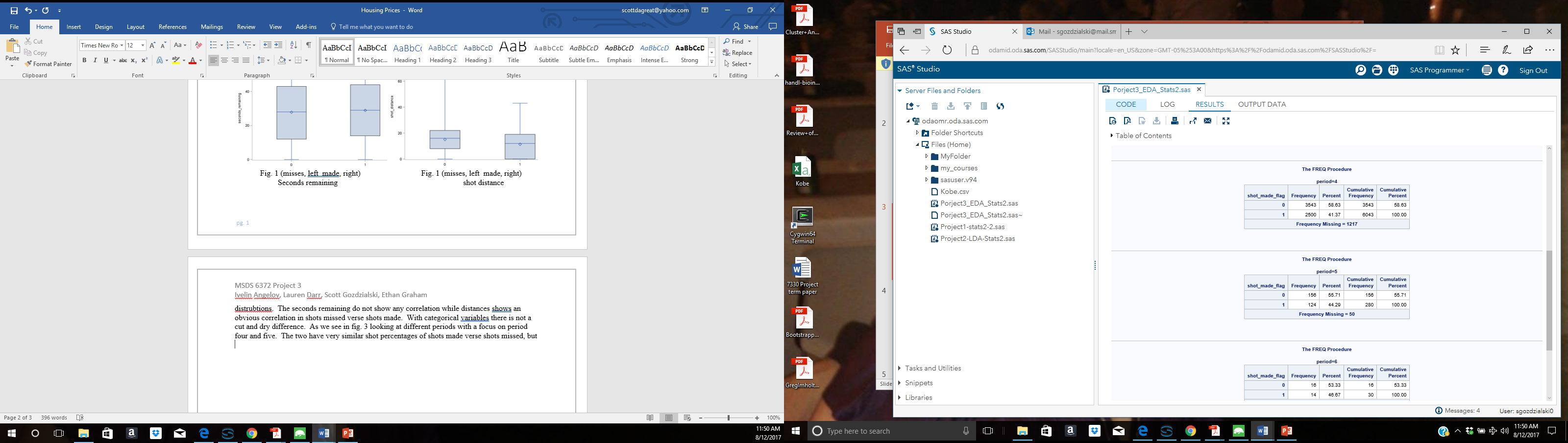
The seconds remaining do not show any correlation while distances shows an obvious correlation in shots missed verse shots made. With categorical variables there is not a cut and dry difference. As we see in fig. 3 looking at different periods with a focus on period four and five. The two have very similar shot percentages of shots made verse shots missed, but with different standard distributions we end up with completely different statistics. With the greater count of shots taken in period four than five the small percentage difference makes a greater difference than the difference in in period five.

Fig. 3

# Analysis

# Appendix