**I. Summary**

Our group has taken a study “Graduate Admissions” as performed by Mohan S Acharya and his associates and applied various Big Data concepts we have learned over the semester to the dataset. In this report we will both demonstrate what we have learned in ITCS 6100 Big Data, but also provide insights as to what we have discovered in the process of exploring this dataset.

The study was based on Indian students undertaking Graduate degrees (i.e. Master’s) in American (i.e. USA) Universities. The intended goal was to assist prospective students in creating a short list of universities that would most likely accept them based on seven primary metrics when measured against the odds of Admittance.

1. GRE (Graduate Admission Examination)
2. TOEFL (Test of English as a Foreign Language)
3. University Rating
4. SOP (Statement of Purpose),
5. LOR (Letter of Recommendation)
6. CGPA (Undergraduate GPA)
7. Research experience ( Yes = 1 || No =0)
8. Chance of Admittance (ranges between 0~1)

**II. Insights from Data Visualization**

As part of the process of data processing we generated various visualization of our data, in order to see if there were any immediately recognizable patterns or irregularities present that would otherwise be impossible to ascertain simply from looking at an Excel spreadsheet. This is what we found:

1. No two variables have a negative correlation; this indicates that there is a synergy at work among the metrics.
2. Looking at variable plots, we realized that most Universities could be considered average in terms of their rating.
3. From the correlation table we learned that there appears to be a strong relationship between “Chance to Admit” and “CGPA” There was also a weaker but, noticeable link between SOP & LOR.
4. From the box plot we have surmised that students with Research experience have greater chance to admit.

From these insights we generated our initial 3 Hypotheses:

1. *Strong SOP & LOR scores have a higher chance of getting an admission offer.*
2. *Research experience matters more than any other criteria for Admission.*
3. *University rating has no bearing on whether a student receives an offer for Admission.*

**III. Insights from Modelling**

After generating our initial Hypotheses gleaned from the data visualization, we then constructed various models to a explore the dataset further. We relied on logistic regression, classification trees, and linear regression. Whilst working with the data we proposed two more working hypotheses:

1. *A higher CGPA corresponds with a higher chance to be offered a chance of Admittance.*
2. *GRE predictor provides more insight than TOEFL; TOEFL should be dropped as a predictor.*

After building our models we decided to test out hypotheses against each model as opposed to selecting a single model for each Hypothesis. We felt that this approach would generate more insights, since it would provide more rigorous testing when examining our hypotheses.

**IV Recommendations**

After completing our exploration of “Graduate Admissions” we have garnered some key insights. We will direct these insights toward prospective students, as that was the focus of the preparation of the data sets:

1. SOP & LOR do have an impact, but it is circumferential; they can tip odds in one’s favor, but they shouldn’t be regarded as one’s focus.
2. Having research experience while valuable, isn’t a core factor in getting admittance. It will increase the odds in one’s favor, but it is not a core component.
3. University rating appears to have little to do with a student gaining admissions. The result gained from testing against each model would suggest a more in-depth study should be conducted.
4. The single best thing a prospective student can do to gain Admittance is to develop and maintain a high CGPA.
5. Take the TOEFL and the GRE is redundant; one can surmise that if a student struggles with the English language then they are unlikely to do well on a GRE administered in English; make the TOEFL redundant. However, as one of the models held the hypothesis to be false, we would recommend another more in-depth study be held to validate or refute such a strong claim.