STOR 565 Project Proposal
Group 7

1. **List of members:** Di Qin, Shuming Sun, Ahmet Hatip, Ellie Lan, Scott Smith

2. **Description of Project of Interest:** RentHop is a rental listing company operating in New York City. Our project is to classify each rental listing according to the projected number of inquiries it will receive. Response variable is a categorical variable that has three different levels: low, medium and high.

3. **Description of Data**
   a. **Data Source**
      i. The data source is from a completed Kaggle competition: https://www.kaggle.com/c/two-sigma-connect-rental-listing-inquiries/data
   b. **Data Size**
      i. The data consists of the following files:
      ii. train.json - the training set with size of 67.3 MB, 52730 observations
      iii. test.json - the test set, we will not use it because there is no response variable
      iv. Kaggle-renthop.7z - listing images organized with compress size of 78.5GB
   c. **The variables that will be used in our analysis include the following: 1)** Interest level: the response variable and the one we are trying to predict, categorical with high, medium, low; 2) bathrooms: number of bathrooms; 2) bedrooms: number of bathrooms; 3) building id, listing id, manager id; 4) created: time/date stamp of when the listing was created; 5) description: free text description 6) features: a list of features about this apartment; 7) latitude, longitude; 8) photos of listings; 9) price: in dollars; 10) the street address and display address

4. **Explanation of Types of Machine Learning Techniques**
   We will treat this project as an ordinal classification problem and consider many different techniques described in class including:
   - **Variable selection:** Lasso and ridge, etc
   - **Dimensionality reduction:** PLS and PCR, etc
   - **Classification**: KNN, Logistic Regression, LDA, QDA, Support Vector Machines, Tree methods, etc.
   - We will also need to use natural language processing packages such as spaCy on the description of the listing and deep learning on the image data.

5. **Potential Challenges and Approach to handling these**
   - The largest potential challenges we face are using the free text data and image data as inputs for predicting classes. Images and free text will lead to the extraction of a very large number of features which would overwhelm the model. We would like to approach this by preprocessing the free text and images using various feature selection and dimensionality reduction techniques to select only the features that provide the most predictive power. We would then add only those features to the remaining features for classification.
   - Another challenge is how to use the geographic data (latitude and longitude) as predictors. One approach is to use a weighted KNN approach to consider only 'neighbors' very close to the subject property.