Project Number	Title	Short description/Abstract	Group members
1	**The State of Masculinity**	There has been much evolution of people's conceptions of gender and gender roles in recent years. This dataset contains men's responses on questions related to masculinity. I plan on using the data to find (classify) dominant outlooks and their relevant predictors. To do this, I will use various methods that we have learned so far, especially PCA and LASSO. I was not able to find a group and would be happy to join or accept others who were also not able to or join an existing open group.	Reid Doctor,,,
2	**Will Your Yelp Review be Useful?**	A special feature Yelp provides is community voting on reviews. Users have a choice to vote a review: Useful, Funny, or Cool. It would be very useful if one could predict which reviews will become popular among users before accumulating lots of votes. Fortunately, Yelp has an abundance of data to make this project a possibility. With data containing, businesses, check-in sets, users, and reviews, we aim to predict the number of Useful votes a review will receive.	Guillermo Rios Martinez,Nicholas Martz,,
3	**League of Legends**	We are going to look at data gathered from thousands of games and fit models with several subsets of predictors. Our goal is to figure out which factors, such as destroying towers and buying weapons, are most important for winning. After finding the best model we will use it to make predictions about current games. The data set we are using is from Kaggle, and has 57 predictors for 7620 games. We will be using forward stepwise selection to create models from subsets of the predictors and choose the best as our final model.	Phillip Lagoc,Ronald Tun,Louise Jensen,
4	**Predicting Substance Abuse With Multinomial Logistic Regression**	We are utilizing the "Drug Consumption (quantified) Data Set" from the UCI machine learning repository. With this dataset, we will see if we can accurately classify subjects into levels of drug abuse based on predictors found in the dataset. Predictors in the dataset range from education level and age range to personality traits like impulsiveness and agreeableness. There are 5 classes of drug abuse, ranging from "Never Used" to "Used Last Day" that each person can fall into. Since we are predicting categories based on quantitative data, we figured that we would utilize a multinomial logistic regression model for our data. We are going to fit each parameter step-wise until we can find the best subset for predicting the class of drug abuse. Once we have our best subset of parameters, we are going to check the strength of our model by using k-folds cross validation and bootstrapping.	Dylan Bragdon,Hayden Telson,,
5	**Gun Violence**	We want to analyze gun violence data, and assess which factors influence the frequency of gun violence most. We're looking at a dataset from Kaggle, which includes data that was collected (with additional processing) from the Gun Violence Archive. Some factors we'd like to look at include: how often shootings happen in open-carry states vs non-open-carry states; relationship between frequency of shootings per county and that county's political party or income; the relationship between states that give the NRA the most money and the number of shootings that occur in those states, etc. The analysis approaches we plan to use include clustering, classification (for categories like political party), regression (for continuous values like income), etc. We will use cross validation and bootstrapping in order to verify that the model we choose is a good fit. We will also have many visual representations of our data, such as maps (using latitude and longitude found in our data set) to clearly show any relationships we may find.	Yasmine Nassar,Hiba Dahbour,Noor Dahbour,
6	**Police Calls for Service in San Diego County**	Our topic of choice deals with police calls for service in San Diego County. In particular, we want to address the question: What kinds of police emergency services are prevalent in different subregions (zip codes) of San Diego County? We will utilize the dataset from: https://data.sandiego.gov/datasets/police-calls-for-service/ paying special attention to the variables street, call type, and priority. We will utilize both regression and classification analysis to help answer our question. We will be clustering the crime codes to its respective incident location (zip codes). Something we want to explore is if we can predict if a certain crime will happen based on a given location. For example, after classifying our data we see that mostly public intoxication infractions happen on Garnet street, we can utilize these observations to see if the next crime reported at Garnet street will be public intoxication. This particular sub problem will utilize regression analysis.	Anuraj Dash,Syed Zain Ali Baquar,,

Project Number	Title	Short description/Abstract	Group members
7	**Breast Cancer Diagnosis based on the Physical Features of the Tumors**	As the risk of female's exposure to breast cancer increases, it is important for us to synthesize the data we can reach and find an efficient method to determine the property of patients' tumors in the purpose of predicting patients' risk of having breast cancer at a premature stage. We will use the statistical summary for the measurements of characteristics of every patient's cell nuclei, which is calculated based on the digital image of patient's breast. The dataset we are using is the Breast Cancer Wisconsin (Diagnostic) Data Set. This data set contains ten consequential features of the cell nucleus for the diagnosis: radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry and fractal dimension. And for each feature, the data set includes the mean, standard error and the worst measure. To do the analysis, we will use logistic regression for the classification of "malignant" or "benign" tumors, cross validation and LASSO as well as Ridge regression for model selection.	Zhiran Chen,Wenxi Xu,,
8	**Gene expression analysis**	Knowing that different gene expressions can lead to different proteins, which can in turn influence how we produce and excrete insulin, we believe there are specific genes involved in this process. We are searching for datasets containing cell information for both diabetic and non-diabetic people. So far, the data we have found has a much greater number of features that samples. Therefore, we hope to perform dimensionality reduction before eventually fitting a linear (or logistic) regression to the data.	Jonathan Wells,Zachariah Gutierrez,Aaron Liu,Aaron Wong
9	**Relationship between street condition and the number of traffic collisions in city of San Diego**	We use datasets from San Diego government. Street Streets Overall Condition Index (OCI) including the City performs street condition surveys and assigns an Overall Condition Index (OCI) number to each street. OCI is only one of many factors to determine the order of street repairs, and Traffic collisions set including traffic collision reports within the City of San Diego. Generally a report is not taken for property damage-only collisions that do not involve hit & run or DUI. According to our search, bad road conditions are critical reason of car accidents. Even though the drivers are good at driving, they are required quick reactions to deal with the potential hazard on the road. Therefore, we want to conduct this research in order to find the correlation. We think that if the street condition is bad, drivers will face more difficulties on driving. Therefore, we predict that the worst street condition will increase the number of traffic collisions. By clean and manipulate the two dataset, we can merge them together based on the street name, the coloums can be "street name", "collision reports amount", "amount of injured/death ppl" and "overall condition". And then to fit them into multiple models to find the relationship between them.	Haoxiang Yu,Hongming Zhang,Ruisheng Wu,
~~10~~		*(Project 10 was a duplicate; do not comment on this one)*	
11	**Stock Portfolio Optimizer**	In this project, we will attempt to make an optimal stock portfolio that maximizes potential returns while minimizing risk. We will accomplish this by clustering stocks by their Sharpe Ratio, which is a measure that compares past returns to past volatility (a proxy for risk), then selecting the stocks from the top cluster. We will also use a regression to predict which stocks will perform well and select the top performers. We plan to use data on stocks from the S&P 500, which is a collection of the 500 largest companies in the United States. We will get our data from this URL: https://www.kaggle.com/camnugent/sandp500 . We hope that this project is useful to young people who want to get started investing, but are unsure of themselves.	Neal Vaghasia,Sahil Patel ,Cole Reynolds,
12	**Analysis of Storm Fatalities and Variables Contribution to Indirect or Direct Fatalities**	We propose using storm location and fatality data from March to June of 2018 gathered throughout multiple states in the United States to analyze which types of accidents were not only most common, but also yielded the most fatalities. By using a multiple linear regression model with variables such as time of the year the incident occurred as well as the type of fatality, the resulting regression could reveal which of these variables are greatest influences to whether or not a victim suffers from a direct or indirect fatality. The data set we will be using to create models will be gathered from the National Centers for Environmental Information. The model that is produced from our analyses can be used to predict what accidents victims are most likely to experience during a certain time period/month, as well as which ones would result in the highest chance of a direct fatality as opposed to an indirect fatality. Furthermore, such a model could also help natural disaster response teams to be better prepared to minimize fatalities in case of impending storms.	Niki Tam,Arshia Sehgal,,

Project Number	Title	Short description/Abstract	Group members
13	**Crime in Los Angeles**	We plan on investigating crime rates in the city of Los Angeles. We will look at different predictors, such as gender, age, and type of crime. The dataset we have collected, has all of these predictors as well as others that we can potentially look at. In our data, we have categorical variables with more than one level, so we plan on using analysis techniques such as logistical regression. We also plan on using either cross validation and/or bootstrapping in order to determine how likely someone is to commit a crime, or how likely a crime is to be committed on a certain individual. We also plan on using clustering, in order to see the likeliness of these factors occurring. We also plan on possibly looking for other data, and merging the two datasets in order to investigate further possible correlations and predictions.	Kelvin Murillo,Karl Chen,,
14	**Kaggle TED Talks Dataset Exploration**	The TED Talks dataset contains 2550 data, each with 17 attributes, including numerical attributes such as date posted, video duration and ratings, nominal data such as tags, and textual data such as name and transcripts. We plan to explore trends in TED topics and other interesting patterns. While we have not decided on the specific tasks to study, possible options are as following: (a) Use numerical data to predict each other, potential models include different forms of regression; (b) Convert textual data such as transcripts of each video into a document vector (using statistical models, word2vec, and so on), perform unsupervised clustering to observe interesting results; (c) Use the document vectors to predict tags/ratings, potential models include different logistic regressions and neural networks; and so on.	Jingya Huang,Wenlong Zhao,Yuezhou Sun,Zhaokai Xu
15	**Gene related to diseases**	We got the gene data expression that linked to different diseases. And we wanted to see if a certain kind of gene expression is related to a disease. We will use logistic regression to train the model. If the probability is bigger than 0.5, we predict it as positive to the disease.	Qingying Luo,Hongyi Pan,,
16	**Using News to Predict Stock Movements**	We use the content of news analytics to predict stock price performance to help investors to make better decisions. By analysing the news data to predict the stock movement, we can harness this power to generate a great impact all over the world. We will do EDA (Exploratory Data Analysis) first and get a brief understanding about the data. After choosing the features, we will try some popular methods to do modeling, such as KNN, linear model, SVM, random forest, etc. During this procedure we will also continue to analyse the data feature for better performance. Finally, we will explore good methods for ensembling models such as bagging, boosting or stacking. Data for this project comes from the following sources: Market data provided by Intrinio. News data provided by Thomson Reuters. Copyright ©, Thomson Reuters, 2017. All Rights Reserved.	Linsong Wang,Tianyu Sun,Yifu Zhou,Ziqi Wu
17	**Science for Money**	We want to see what factors play the largest role in research receiving funding and the amount they receive. Look at what predictors lead to success in attaining research grants such as Google Scholar references, google trends popularity, institution, geographical region, grant amount, keywords. We will perform a regression analysis containing these predictors and utilize cross-validation on grant data to find the optimal model. Once we have a model, we will provide a simple web interface for allowing users to enter their own predictor values to see how much funding a specific research project will most likely receive. Overall, we are interested in seeing what correlations	Luyanda Mdanda,Kaylani Kottilil,Ian Carrasco,Jesse Kim
18	**How to get high rating on Google Play Store**	We will analyze the dataset of Google Play Store Apps and hopefully we could predict an app's rating based on its category, price, number of downloads, etc. We chose our dataset from Kaggle, and the dataset is uploaded by Lavanya Gupta. We may use linear regressions, forward and stepwise model selection, k-fold cross-validation, etc. during the analysis.	Ruxuan Ma,Zhang Zhang,,

Project Number	Title	Short description/Abstract	Group members
19	**Measuring Crime Probability and the Potential Regulation of Crime**	Out of 117 countries, the United States ranks at 47 for having the worst crime rate of 47.01% -- Venezuela has the highest rate of crime with 82.38% and Japan has the lowest rate of crime with 12.69%. For our group project, we plan on measuring the probability of specific crimes that may continue to occur in the United States based on previous crime rate data found on Datazar. Within this dataset, we have 19 observations with over two hundred million people in a population and over 1 million people that commit a violent crime during a specific year. This shows us that even though we may run into problems of having trouble of regulating crime with potential noise factors of gender and location, we still have a large dataset to work with. This large dataset can still give us statistically significant information that will help us with our project. Therefore, with this information, our group then plans on using the predictions we make through modeling and data analysis in order to come to a potential solution that can regulate crime more. By using modeling and data analysis, we plan on using some or all of the following statistical models: logistic regression, box plots, scatter plots, clustering, and k-nearest neighbors. These statistical models can be applied to our project in terms of helping with prediction.	Darin Lee,Kenneth Truong,Caitlyn Gonzales,Stella Khachatryan
20	**Analyzing Popular Food Trends and What categories make the best business**	Our topic is about the popularity of restaurants and the food trends associated with them. We will analyze the yelp dataset which has restaurant info, customer info, and review data. We will try to use both classification models and regression models to draw conclusions from the dataset. Some ideas we have are predicting how many stars a restaurant will receive based on it's characteristics. Another is using clustering to determine what type of restaurants do well on yelp, and what type do not. We can also compare average stars across cities to see what city has the best food so you can plan your next trip accordingly.	Kevin Rafferty,Pete Sheurpukdi,,
21	**The Factors that Affect Winning a Professional Tennis Match**	We are interested in taking data from https://github.com/JeffSackmann/tennis_atp and analyzing it in order to find out the effect of various factors such as player attributes (height, weight, age, ranking etc) as well as court conditions (surface, indoor/outdoor) and match statistics (aces, break points, winners etc.) on winning. We intend on analyzing the impact of these factors, as well as how these factors have changed over a 5 year span of tournament results. The data that we plan on using is a compiled chart of matches in all the tournaments from the professional ATP (Association of Tennis Professional's) men's tennis from the years 2012-2017. Using all of the predictors at our disposal given by the data, we would first want to get a sense of how the data is distributed which could be done by a basic scatterplot. From this we can start to investigate potential relationships in the data. The basic framework would be what is the probability that a player will win given β1, β2,...βN. Essentially, this is a classification problem, so using techniques like LDA and K-nearest neighbors, we can start fitting models given our data points.	Ben Cauffman,Jeffrey Lee,Ruyin Zhang, Henry Huynh/ A11687994
22	**Age vs Drug Use Dependency**	Does early age use of drugs affect later drug/alcohol dependency. Data sources may include lab data sets from one group member who works at a lab, previous government drug addiction studies, and online data sources. Our analysis approach may mostly be focus on linear or multiple-linear regression using a scatter plot to show if there are any statistical significance on the variables.	Henry Wu Ou,Cameron Lee,Maria Bordyug,
23	**Examining the relationship between student loan debt and completion, repayment, and career earnings.**	An extremely troubling concern for many post graduates out of college is paying off their student loans. Many people take out big loans in hopes of recovering the costs with their future careers, only to be met with an unstable job market and lackluster salaries. I plan to analyze the relationships between student loan debt and repayment, degree completion, career earnings; essentially wanting to answer whether in general, is it worth it to take out loans for one's college education and to what extent/limit. I plan to use the data sets provided on https://collegescorecard.ed.gov/data/ and possibly others that I find later on. Briefly looking through the datasets, I am planning to use regression to analyze the relationship between the various features.	Shirley Yu,I wasn't able to find any partners/team to join at this time. :(,,

Project Number	Title	Short description/Abstract	Group members
24	**Identifying the correlation between meteorite landings and natural disasters on Earth**	No recording of human fatality is attributed to meteorite landings in the past one thousand years, as reported by National Aeronautics and Space Administration's (NASA) Jet Propulsion Laboratory. However, meteorite landings can create impact events that have hazardous effects on the environment and subsequently cause natural disasters that lead to thousands of deaths of human and animals alike. Therefore, we are interested in examining the correlation and possible factors attributing to the relationship between meteorite impacts and number of subsequent natural disasters such as earthquakes, tsunami, and forest fires. We hypothesize there is a significant relationship where more instances of natural disasters occurring after the impact events, and specific attributes of the meteorites are positively correlated number of natural disasters that follows. We are using Meteorite Landings data from NASA's Open Data Portal and Disaster Declarations from Federal Emergency Management Agency (FEMA) to address our problem. First, we use nonlinear regression to consider whether a relationship exist or not and identifying, if any, which aspect of the meteorite attribute to a significant correlation to the number of natural disasters. Then, we use bootstrap sampling to estimate that distribution of our model coefficients empirically and cross validation to identify the accuracy of our model. The correlation and attributes identified in the study will enable further understanding of the impact meteorites have on Earth.	Lauren Liao,Jiwon Yoo,,
25	**Mental Health in the Workplace**	We are seeking to identify any factors in a workplace that may contribute to mental health issues. These factors may include amount of coworkers, mental health benefits provided, or the nature of the work that they do. The dataset we are using samples over 1,000 people and asks questions about different qualities of their workplace, the attitude of their bosses and coworkers surrounding mental health, and if they have every been treated for a mental illness. This is important to those already struggling with mental health issues in identifying work environments that won't negatively affect their mental health. We plan on using multiple linear regression in order to find out if there is a significant effect of any predictor, and which predictors are significant.	Avery Jones,Ryan Lee,Eric Woolsey,Ty Garrett
26	**The Effect of Student Programs on Graduation Rates and Student Success**	The project explores factors contributing to K-12 student graduation success. A multiple logistic regression model will be used to look at the relationship between dropout rate with SAT scores, high school graduation rates, private vs public teacher salaries, and student to teacher ratios. The Head Start federal program will also be examined for its effectiveness in bridging the education gap for lower socioeconomically populations. Data will be taken from government sites including the National Center for Education Statistics (NCES) and the Department of Health and Human Services. In order to parse through the data and make accurate predictions on data not used during our trials, we will also utilize cross validation methods.	Shreya Sheel,Nicole Askar,Kian Falah,
27	**Does gender predict the cost of healthcare?**	The topic we are interested in is predicting whether someone's gender can predict their healthcare cost. We are using a data set from the US Department for Health and Human Services titled "Compilation of State Data on the Affordable Care Act". The data set has 50 data points for each variable, and has over 70 categorical and numerical variables. We are focusing on two variables, "Adult Males with Lifetime Limits on health benefits pre-ACA" and "Adult Females with Lifetime Limits on health benefits pre-ACA". We are planning on using classification to create a model predicting healthcare cost with gender as our predictor. Then we plan to use K-Nearest Neighbors classification to compare models and find the best fit classification boundary.	Abby Kostukovsky,Ruby Tazim,Justine Hormigas,n/a
28	**Analysis and prediction of malignant tumor vs. benign tumor**	This project we focus on breast cancer, which is the most common disease in women. Our main goal towards this project is the analysis and prediction of malignant tumor and benign tumor. Our dataset is from open-source dataset website (https://www.kaggle.com/uciml/breast-cancer-wisconsin-data?), so there is no ethics concern. The dataset contains 570 observations of breast cancer patients in Wisconsin, and there are 30 features of breast cancer and diagnosis of whether it is malignant or benign. We will perform PCA and stepwise selection to reduce the dimensionality of the dataset while preserving the accuracy of our computation. We will use logistic regression, LDA, SVM to perform cancer types classification. We will evaluate each model by performing cross validation using grid search and bootstrap.	Lei Wan ,Zeyun Wu ,Zhenxian Lu, Yundong Wang

Project Number	Title	Short description/Abstract	Group members
29	**Airbnb Pricing Predictions**	The problem we observed Our topic is to predict the prices of Airbnb listings in San Diego by neighborhood, number of rooms, and availability. We hope our findings will facilitate the process of finding affordable Airbnb listings. Our data sources are the public data sets disclosed by Airbnb. The data is compiled from input given by owners of Airbnb listings. We will use non-linear regression to visually portray the correlations.	Boya Ouyang,Elimelec Pineda,Won Suh Kim,Haimei Yu
30	**LoL is Not a Joke**	With the League of Legends World Championship just recently ending, players from around the world are getting ready for a new season to perfect and hone their skills. The issue here is that each year, competitive players are spending countless of hours trying to perfect strategies that would lead to victory, but aside from individual skill, how can they increase their chances of winning? Our team plans to analyze data sets of League of Legends on Kaggle.com concerning character picks, bans, gold earned, winning team, and other things related to the specifications of a match. These data sets are both quantitative and categorical, so we'll be using various types of models for the different data types accordingly. Whether it's just for fun or playing competitively, our team hopes to answer the question of what's the most effective and reliable way to win games in League of Legends.	Brandon Lien,Sebastian Kleinerman, Franklin Moirao,
31	**Coffee Quality**	For our project, we will look at coffee quality ratings of arabica coffees, and the growing conditions of the beans from which these types of coffee come from. We will focus on making inferences about aroma and flavor qualities, and use country of origin and growing altitude as our predictors. We will use the data from the “Coffee Quality Institute Database”, which can be found here: https://github.com/jldbc/coffee-quality-database. Possible approaches to the dataset may include linear regression models (including dummy variables to use country of origin as a predictor) and cross validation/resampling methods such as kfold, LOOCV, and the boostrap. We may not be using classification models as the dataset does not include categorical data results.	Nathan Chau,Jonathan Fong,,
32		We will be looking at data that outlines fentanyl and cocaine overdose deaths. The data has information about rates of overdose deaths based on state and ethnicity, along with other parameters. Our goal would be to find areas with higher rates of overdose deaths, which would be the best areas to divert resources, should the federal government ever choose to tackle this problem with policies. We would try different models, including linear, logistic, and k-nn, using cross-validation to compare and ultimately choose the best model for this dataset.	Tyler Chau,Sahba Mobini Farahani,Brigid Overton,
33	**The Perfect Formula to a Popular Boba Shop**	https://docs.google.com/document/d/1iS423K7kq3asE9eYbjr09AnKLHdIi9qkBqGM8xXhbH8/edit?usp=sharing Boba or bubble tea is a real popular drink among students at UCSD. We wanted to get into the business analytics of Boba shops and discover the factors that may point towards their popularity, such as number of reviews and gimmicks/attractions- that we can extract from the reviews using certain keywords. We will be measuring popularity of the place in terms of star-ratings weighted by the number of ratings. Data Sources: Yelp DataSets: Yelp has relevant dataset on https://www.yelp.com/dataset	Sowmya Parthiban,Cody Smith,Niki Tran, Matthew Montehermoso
34	**Classifying Schizophrenia patients from EEG**	Using pre-processed EEG data from 14 healthy and 14 schizophrenia (SZ) individuals. Based on differences in the EEG from these two groups, we will attempt to create a logistic classifier that predicts the probability that an individual has SZ based on their EEG. We may also implement some regularization given that there are data from 16 electrodes, and various frequencies, and presumably not all of them are "important". Some concerns we have off the bat are 1) the size of our EEG files are very large and 2) we are not quite sure what dimensions of the data we should use (e.g. Time vs. Frequency band, or Time vs. Amplitude, or should we average each person's EEG across electrodes, thus losing information about the electrodes, etc.)	Rebecca Eliscu,Mikey Malina,,

Project Number	Title	Short description/Abstract	Group members
35	**World Happiness**	In our project, we will explore worldwide happiness and the extent to which different factors play a role in the happiness of individuals. We will use a dataset titled “World Happiness Report” taken from kaggle.com (https://www.kaggle.com/unsdsn/world-happiness/version/2). The World Happiness Report is described as a “landmark survey for the state of global happiness”. It contains data on 155 countries, ranking each country on its happiness score based on predictors such as economic production, social support, life expectancy, freedom, absence of corruption, and generosity. The happiness score was obtained by using data from the Gallup World Poll, which asks a question called the Cantril ladder. The Cantril ladder “asks respondents to think of a ladder with the best possible life for them being a 10 and the worst possible life being a 0 and to rate their own current lives on that scale”. We are interested in seeing if countries from the same region have similar happiness scores. We will analyze this by making a scatter plot of happiness score vs. happiness rank and then coloring each region (Western Europe, North America, Australia and New Zealand, etc.) differently so we can see if the regions cluster together on the scatter plot. We are also interested in seeing the extent to which economy, health, trust, and generosity play a role in the happiness scores of each country. We will perform a single linear regression using economy as a predictor, then perform a multiple linear regression using the other predictors to see if economy is still statistically significant, and if any of the other predictors are statistically significant.	Dustin Crotty,Emi Oda,Ezekiel Samatua,
36	**The relationship between purchasing quantity and price of the avocado**	Avocados has been rising in popularity across the country over the last couple of years, probably due to its increased circulation around various social media sites. With such an accumulated demand, we want to explore whether there is a correlation between purchasing behavior and the price of an avocado. With that, we can further explore the trends of regional purchasing behavior as well as whether or not more people are buying conventional versus organic avocados. The dataset we will be using is called “Avocado Prices” which was found on Kaggle. This dataset looks at different retail markets around the United States in the avocado’s sales volume and price from 2013 to 2018. We will be analyzing this data by applying clustering when looking at purchasing patterns in different geographical locations and regression (linear and nonlinear) models when considering how prices are impacted by year.	Krislyn LaCroix,Judy Tang,Jordynn Bartolome,
37	**How to Become an NBA Superstar**	We plan to use player stats within a constrained number of years to predict if young players who are in the early stages of their careers during this time period will later on become an NBA superstar. We will find all player data from nba.com and basketball-reference.com for seasons 2013-current so we can make sure that these predictions will be applicable to the modern NBA. These sites contain all player statistics since the 70's, which is when the NBA began to recognize and record all of the statistical categories we know today. We will define the term 'NBA superstar' by averaging the career statistical outputs of the players in today's NBA who are widely recognized as superstars by NBA. This will create a threshold that will be used in our multiple logistic regression to be met or passed by current young players. This threshold will contain these statistics: points per game, rebounds per game, assists per game, steals per game, average field goal percentage, average free throw percentage, steals per game, blocks per game, 3 pointers made per game, and average win-shares (a statistic that measures an individual's average contribution to his or her team's victories). These categories will become our predictors in our multiple linear regression. Then we will see if certain young players from years 1-4 in their career can be expected to meet or surpass at least 4 categories our 'NBA superstar statistical threshold' in future seasons, which marks their entrance into superstar when achieved. The reason they only need to meet or surpass 4 of these average category values is because average values for certain categories fluctuate greatly between different positions (point guard have higher assists and steals, centers have higher rebounds and blocks, etc), and therefore these current superstars have been seen to outdo each other in 4 different categories when they play different positions.	Hao-in Choi ,Raju Ivaturi,Ye Lin,

Project Number	Title	Short description/Abstract	Group members
38	**Breast Cancer Prediction**	We find data on UCI machine learning site about breast caner consisted of two parts. The data sets contain same 30 predictors, numeric features such as the radius and perimeter of the cell nucleus, and ID and Outcome data. However, one of the data sets' Outcome column contains actual Diagnostic result, which stands for real observations, while the other data set contain Prognostic result. which is predicted through empirical method. It is also noteworthy that the Prognostic data set contain 2 extra predictors, Tumor size and Lymph node status. The Diagnostic data set contain 569 observations, and the Prognostic data set has 198. We plan to fit a model using the Diagnostic data with 30 predictors and then fit another model based on the Prognostic with same 30 predictors but also Tumor size and Lymph node status. By applying two models with different predictors used on the Prognostic data set, we plan to examine the superiority of the models, and the weight of influence for Tumor size and Lymph node status on formulating Outcome prediction.	Yurui Feng,Samantha Sze,James Taniguchi,Risheng Tan
39	**What makes the best restaurant rating?**	We want to answer the question, what makes the best restaurant rating? To do so, we want to the explore the relationship between the type of the restaurant and its ratings. We will analyze whether certain categories or combination of categories can predict the success of the restaurant. Our analysis approach is to find a set of common categories that restaurants share, such as location, food type, opening hours, etc. and use them to create a model to predict the rating a restaurant. We are using Yelp Dataset to conduct this analysis. More specifically, we are using business.json as of 11/15/2018.	Timothy Lue,Simon Li,,
40	**Predicting Wildfire Properties using Data Models**	With the rate of wildfires increasing recently, it's important to be able to predict their properties. This way firefighters and other aid agencies can efficiently allocate resources to contain fires. The first data set we'll use contains meteorological data (humidity, temperature, etc.) taken during wildfires in a natural park in Portugal and records the area of these fires. The second set is a New York state wildfire reporting database that also records fire size but contains different features such as fuel type and cause. The third and final set is a NASA fire data set that records the brightness intensity of fire incidents and its coordinates. Through each data set, we hope to make models that can receive features such as temperature or coordinates and predict fire size or intensity. Because there are multiple predictors, multiple linear regression will be used while cross validation will be used to estimate MSE and model quality. To find out which features form the best predictor model, stepwise selection may be used as well. If there are abnormally highly correlated predictors, ridge regression will be used to eliminate them and make our estimates more accurate.	Omar Ahsan,Tatiana Goodwin,Bailey Bartley,Jason Monroy
41	**Predicting Wine Quality**	We want to predict the quality of a wine based on different factors, including fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol. We will first fit the data using both stepwise and ridge regression to get the "best" parameters. We will then compare the two models using 10 fold cross validation. Once we get the better model, we will fit all the data to that model and get our parameter estimations. Finally, we will use bootstrap resampling to determine error bars for our model parameters.	Lillian Lee,Noah Carniglia,Albert Carillo ,
42	**The Relationships within Relationships**	Human relationships are a topic with hundreds of books written on it, but surprisingly few scientific studies on the issue. Using hundreds of responses from online surveys, lots of practical insights into the nature of relationships can be made. For instance, we hypothesize that females tend to develop higher level of attachment to their romantic partners compared to males. To investigate whether gender impacts the attachment level we are going to use a linear regression model. We also suspect that age might be another factor in a person's attachment to their significant others, with older individuals feeling more attached than younger individuals. To test this, we will also incorporate age as another predictor variable to analyze its effect on attachment level and whether or not there is an interaction effect between the two predictors. Our data source is from an online public personality test, which informed responders that their data would be used for research. The database contains various questionnaires, but for our focus, we chose the survey on "Experiences in Close Relationships Scale." (March 2018) Our dataset: Experiences in Close Relationship Scale (3/1/2018)	Reginald Uy,Mingbin Li,Natalyia Ratosh, Lindy Wong

Project Number	Title	Short description/Abstract	Group members
43	**Parking in San Diego**	For our project we plan on looking into the parking issue in downtown San Diego. There is a severe lack of parking in the downtown San Diego area, and with the growing population the problem will only get worse. We want to see if we can analyze data based on parking meter locations, popularity, and locations of local business and use this to find the most crowded parts of downtown and identify areas that need more support. We found two data sets on parking meter locations and parking meter transactions that we can use for our project. We plan to implement K nearest neighbors to find clustering of popular parking spots and multiple regression to analyze our data.	Veronica Merrill,Adrian Dinescu,,
44	**NBA Player Position Classification**	We would like to use data from NBA statistics. With this we would like to perform a classifications on basketball player's positions. In other words, trying to see if it is possible to determine a basketball player's position (Point guard, Shooting guard, Small forward, Power forward, Center) with that one player's stats. We are planning on using the machine learning algorithm such as multiple logistic regression to identify the player's categories. Since logistic regression classifier is mostly used for binary classification problems (problems with two class values), we will go a step further and use multiple logistic regression since there are at least 5 types (5 basketball positions) of classifications we are going to have to do.	Rodrigo Efraim, Gurpreet Dhillon

Project Number	Title	Short description/Abstract	Group members

Project Number	Title	Short description/Abstract	Group members