Assignment 2

Anish Narsian A98071125
Dhruv Kaushal A10794472

1. Dataset

The dataset that we are using for this project is of freshman admits taken in by the University of California from different cities across the state. The dataset consisted of the city, number of applicants from the city and number of enrollees for every year from 2010 to 2014. We used an additional dataset to get the city population statistics which consisted of the city name and population, from the US census. We also had a dataset for the property prices in cities of California which consisted of the city name, median house rate, property price by square foot and rent per square foot, from Zillow. And our final dataset consisted of California unemployment rates, where the table consisted of city names, and unemployment percentage and workforce size for every year from 2010 to 2014 from the Employment Development Department, CA.
Some interesting things about the dataset- nearly 6,000 people applied to the University of California from Los Angeles alone in 2014, out of which nearly 1,600 enrolled into the system, the highest among any other city on both accounts. La Jolla, on the other hand, averages around 400 applicants and 100 enrollments.

Our data was downloaded from following locations:

a. Primary University of California data:
   http://universityofcalifornia.edu/infocenter/admissions-source-school
b. City population data:
   http://www.city-data.com/city/California.html
c. Property price:
   http://www.zillow.com/ca/home-values/
d. Unemployment Statistics:
   http://www.labormarketinfo.edd.ca.gov/CES/Labor_Force_Unemployment_Data_for_Cities_and_Census_Areas.html

Sample of our primary dataset and what the values in there looked like:

CA High School	County	City	Applicants	Admits	Enrollees
A B MILLER HIGH SCHOOL	San Bernardino	Fontana	61	44	31
ABRAHAM LINCOLN HIGH SCHOOL	Los Angeles	Los Angeles	103	74	55
ABRAHAM LINCOLN HIGH SCHOOL	San Francisco	San Francisco	189	132	98
ABRAHAM LINCOLN HIGH SCHOOL	Santa Clara	San Jose	60	46	16
ACAD FOR ACADEMIC EXCELLENCE	San Bernardino	Apple Valley	9	6	4

Assignment 2

Anish Narsian A98071125
Dhruv Kaushal A10794472

ACADEMIA AVANCE CHARTER SCHOOL	Los Angeles	Los Angeles	7	6	5
ACADEMIC LEADERSHIP COMMUNITY	Los Angeles	Los Angeles	27	17	10

Sample of the City Population Data:

Acton	7596
Adelanto	31239
AgouraHills	20657
Alameda	75641
Alamo	14570
Albany	18969

Sample of Property Price data:

Region Name	Avg Value of Home	Value psq	Rent psq
California	449500	294	1.46
Adelanto	0	0	0.75
Agoura Hills	783500	355	1.69
Alameda	828600	521	2.09
Alamo	1599100	532	2.05
Albany	831400	670	2.65
Alhambra	559200	399	1.68
Aliso Viejo	538400	361	1.84
Alondra Park	520700	420	1.98

Sample of Unemployment Data:

City	Force '14	Unemp '14	Force '13	Unemp '13
Alameda County	812000	0.059	783100	0.074
Alameda city	41000	0.051	41300	0.05
Albany city	9800	0.038	9400	0.031
Ashland CDP	10600	0.07	10600	0.086
Berkeley city	60800	0.047	60600	0.07
Castro Valley CDP	32300	0.052	32000	0.043
Cherryland CDP	7000	0.078	6800	0.109
Dublin city	25000	0.037	16000	0.045

For all of our computation in this assignment we linked data across the cities by storing city data in Dictionaries and via this linking did all of our regression and then on our predictions.

Assignment 2

Anish Narsian A98071125
Dhruv Kaushal A10794472

2. Predictive Task:

- ***The task***

Using the primary data set with the above data we will predict the number of students that will go on to join a UC campus, given a city name and the total number of applicants from the city.

- ***Training, Validation and Baseline***

We will use 4 years of school data (2010-2013) for training, and the most recent year's will be used for validation (2014). We intend to use different models of Linear Regression. Our baseline was using just one year's data, of 2013 to train and of 2014 to validate. The MAE was 6.5. We used the Absolute Error and Mean Absolute error values to determine our performance and validate that our model was good enough/performing within expected or acceptable ranges. The AE and MAE were also used to draw comparisons between our models.

- ***Features used***

The features were:

  - Number of students in a given high school
  - Total Population of the city of the high school
  - Average Property rate in the city
  - Size of work force in the city
  - Unemployment in the city

We felt that this data was relevant for the reasons below:

  - Number of students in a high school, obviously is important as it directly influences who can possibly go to university.
  - Total population is important because this value combined with work force and unemployment determines dependence on the work force, and hence financial ability to go to a University of California. It also accounts for social demographics such as given a large total population, social convention of many people working might drive more people into going to a good school. On the other hand, a small population might drive students to work locally and do local work(Eg. Farm yard work, taking over small business' etc).
  - Average Property rate is primarily seen as super important to the affluence of a city as well as and more importantly driven by the presence of good school districts. For example, Fremont, CA: even though it doesn't have very many top companies in its area, has many good schools which drives property rates very high. This will also account for the fact that high property rates imply mortgages and so it might reduce financial ability to send a student to school.
  - Size of work force: As mentioned earlier in the reasoning, this determines ability to send someone to school based on amount of money in family, and dependence on this money based on total population.

- Unemployment rate: During periods of high unemployment, students might be of the belief that going to school is necessary, moreover high unemployment also reduces ability to send a student to school. Hence regression will account for these parameters and generate a good relationship as required.

The aforementioned were some of the reasoning behind us picking our features, while some of reasoning can direct the model in both positive and negative ways, the regression can account for the overarching resultant of these parameters.

- ***Data Processing***

The data processing we did included serializing some of our city-wise data to generate city –wise features. We also had to do a lot of processing to ensure that the data was parsable in python: we had to eliminate blank spaces in the data, eliminate extra data and also convert data to readable (MS_DOS) CSV to have python read it in.

3. Model

- ***Model used and Justification:***

The model that we propose consists of multiple linear regression models trained on a different year's data and then the final model took the average of the theta values of all the regression models. Each model consists of the following features-

a. Applicants from a city for the specified year
b. City population for the specified year
c. Property prices in city for the specified year
d. Size of workforce in city for the specified year
e. Unemployment rate for the specified year

The listed features were used to predict the number of enrollees into University of California given the number of applicants, population, affluence, workforce and unemployment rates. The reason this model was used by us was because there is a strong correlation between rising unemployment and university applicants and enrollments. The property price were included such that more affluent neighborhoods and cities had a higher probability of admitting.

We averaged Theta values obtained by training our model over the years (2010 to 2013) and then used that to generate an MAE.

The logical basis and reasoning behind why this model made sense is listed above in the *Features Used* portion of the **Predictive Task.**

- ***Issue with Scalability and overfitting:***

We tried using multiple dimensions (x^2/x^3) in the data, it seems that even though this generated a lower MAE, it was over fitting the curve of similarity. We decided to go the way of

linear/singly dimensional fitting variables(y = bx + c) rather than multi-dimensional(y = bx^2+ cX + d).

One major issue with scalability was the fact that we had to obtain statistically significant data from multiple sources, getting relevant data over all the years 2010-2014 for the training meant that the scaling required lot of data collection.

- ***Other Models, issues and decision points:***

A major decision point that we had to face was to decide how many years of data to use and what the significance was. This became a major issue for us for the reason that we were getting MAE values of 12 for our overall model with averaged Thetas but lower in baseline 6.5. We realized that in terms of property rates and unemployment 2013 and 2014 were similarish years. The recession of 2008 skewed the unemployment and property rates causing our model to perform worse if we included those year, which led us to decide to use the 2010 to 2014 data. While the recession had influenced unemployment and property rates, it also changed the thought patterns of society wherein higher unemployment implied people wanted to go to school more. Thus there was a skew in our 2014 predictions based on our overall model. On the other hand using training set as that from the years 2010, 2012, 2013, 2014 and using our validation set as 2011, generated a much better outcome. It came out to be an MAE of ~7, while the baseline in that scenario did worse than 7. We felt that overall in terms of our predictor working over all possible socio economic scenarios it might be better for us to take the average even though it might lead to worse performance in the bad case. That is to say our general case performance will be better than our extreme case performance: which is standard and expected of many statistical models.

We would also like to mention that we didn't consider all these other data sets and features from City Population Data, Unemployment Statistics, and property values. We took these on gradually and saw an improvement in our performance in terms of MAE over time.

Our Model progressed as follows:

Model 1: Number of Applicants only, this could be considered the upper limit baseline, this was quite a high valued number at MAE. We can beat this performance by a sizeable amount

Model 2: Number of Applicants, Property rates

Model 3: Number of Applicants, Property rates, unemployment percentages

Model 4: Number of Applicants, Property rates, unemployment percentages, total city population, size of workforce

For each of the above models we tried various dimensions for the features, but in the end a flat linear model performed the best.

Model 4 turned out to be our best and it logically made sense as outlined in reasoning from *Features used* in the **Predictive task** section of our report.

Assignment 2

Anish Narsian A98071125
Dhruv Kaushal A10794472

4. Literature:

My team is using multiple existing datasets to find a solution to our predictive task. The datasets include statistics for applicants, admits and enrollees in the University of California across all its campuses. California city population statistics, California housing price statistics from Zillow and California unemployment statistics from the Employment Development Department. Other similar datasets that have been studied in the past which are similar to our dataset include Ivy League admissions predictions using affluence of applicants and multiple studies of economically prosperous households. Multiple state-of-the-art methods were encountered by us while researching about this data, some of them being Time Series Regression models and economic statistical models. Since the studies we found were not directly correlated to our predictor, we were unable to compare the conclusion, but the general consensus from all the models was that increasing affluence resulted in an increase in applications to schools of higher education.

The link to some of our reading is outlined here:
http://www.mathworks.com/help/econ/examples/time-series-regression-i-linear-models.html#zmw57dd0e825

This reading included multivariate regression, time series, model analysis etc. These are commonly used in Economic modeling. We found that it was not necessarily a good fit for our modeling here.

5. Results and Conclusion:

6.

Table of some of our theta Value generated:

Theta '10	Theta '11	Theta '12	Theta '13	Avg Theta
-2.723	-2.159	-1.329	-5.011 E-1	-1.678
4.528 E-1	4.31 E-1	4.062 E-1	3.794 E-1	0.417
-1.147 E-6	-2.158 E-5	-9.98 E-6	-2.92 E-5	-1.54 E-5
-3.656 E-3	-5.043 E-3	-5.438 E-3	-7.782E-3	-5.48 E-3
4.375 E-7	4.152 E-5	1.853 E -5	5.674 E-5	2.93 E-5
7.627 E0	1.071 E1	4.93 E0	6.975E0	7.561

Now the Average Theta value was used with the data values got from the 2014 data to validate and generate the final Absolute Error and Mean Absolute Error.

Absolute Error: 14756.93

Mean Absolute Error: 12.007

Conclusions based upon the Theta construction and Error values:

Here is an image of Ye absolute and Y-pred values:

Assignment 2

Anish Narsian A98071125
Dhruv Kaushal A10794472

```
Y absolute:  30  Y pred:  36.0
Y absolute:  23  Y pred:  19.0
Y absolute:  7  Y pred:  3.0
Y absolute:  10  Y pred:  10.0
Y absolute:  19  Y pred:  11.0
Y absolute:  31  Y pred:  31.0
```

There are over 1000> listed, so this proposes an output with a margin of error +-12. While these are smaller schools with fewer students going to UC's there are some larger ones with >300 and also some very small ones like 1-5. Our model predicts well for most medium sized schools, but extreme cases such as ones with very few applicants or every large number of applicants are not predicted by the model very well. For example as below:

```
Y absolute:  4  Y pred:  -1.0
Y absolute:  9  Y pred:  9.0
Y absolute:  10  Y pred:  -5.0
Y absolute:  71  Y pred:  79.0
Y absolute:  53  Y pred:  65.0
Y absolute:  35  Y pred:  30.0
Y absolute:  14  Y pred:  14.0
```

We notice that for extremely small values of Y, the predictions are skewed by our model. Through careful analysis of our model, we figure that this is due to the high positive theta associated with the number of applicants.

We believe our model succeeded very well in its predictions considering the amount of data we used up for the model. It is our belief with more relevant datasets, like ethnicity statistics, among other things, we would have been able to fine tune it even more.