Basic Information

Title: KSL AutomoDeals
Names: Chantel Charlebois, Taylor Hansen, Michael Paskett
E-mails: chantel.charlebois@utah.edu, taylor.c.hansen@utah.edu, michael.paskett@utah.edu
UIDS: u1043299, u0642850, u1144000

Background and Motivation

Almost every person in Utah (and some neighboring states) buying a used car will visit KSL Cars classifieds to look for their new wheels. There are few resources for understanding the rough value of a used car, such as Kelly Blue Book (kbb.com), but such services cannot fully integrate the complex auto market of a local area. By storing and analyzing the prices, details, and options for a certain model or class of vehicle, a prospective buyer can evaluate how good the listed price for a vehicle actually is. With such a model, the user can estimate how much a specific car is really worth, and determine if the vehicle is worthy of a test drive.

Project Objectives

Questions:

- How well can we predict the price of a newly-listed car based on the attributes available in an advertisement?
- Which attributes are most influential in determining the vehicle price?
- What areas have the best price of cars?

Aims:

- Create a regression model that will predict the expected price of a car based on several attributes, such as:
    - year, seller type (dealer, private), mileage, color, title (clean/salvaged), transmission type, cylinders, fuel type, number of doors, exterior/interior condition, listing date, page views per day (when a listing has reached 7 days)
- Create a clustered map of “good deals” in different regions

Benefits:
This project could benefit anyone in the market for a used car, helping them to be informed of the potential value of a car they are interested in.

Data

We will be scraping our data from queries of used cars listed on cars.ksl.com. We have read the robots.txt files for both ksl.com and cars.ksl.com to confirm that there are no restrictions for crawling their website. We have also reviewed the terms and conditions and have similarly found no indication for rules on crawling. There used to be an undocumented API for interfacing with KSL (as of four years ago), but it is no longer publicly accessible, so we will be manually scraping with BeautifulSoup.

To avoid consistently using too much bandwidth on their website, we will begin “historical” data collection by saving .html pages over the course of the project so that we can parse them offline. This will form the major basis of our data set against which we can make comparisons for new queries when looking for a good deal on a newly listed used car.

Ethical Considerations

Stakeholders:

- The creators (us)
- The seller
- The prospective buyer
- KSL

Our incentive as creators and prospective buyers is to find good deals without having to manually spend hours searching through KSL for a good deal. For other prospective buyers, the same applies. The sellers have competing interests, as they would like to sell their car for as much as possible. KSL also has a stake in this project, as it makes revenue from ads and from sellers paying for better listings in order to make their vehicle more prominent.
We anticipate that other ethical considerations may arise as the project progresses and details are worked out, but the above seems to encompass our understanding of the ethical issues at the time of proposal.

Data Processing

Each listing page has a fairly consistent format making scraping feasible for the large number of pages we will be analyzing. The quantities we plan to derive from our data have been listed above in the Project Objectives section. When creating a listing, the user is *required* to list the year, VIN, make, model, body style, mileage, title type, asking price, and ZIP code. A timestamp is also associated with each listing. Together, these are the only features we can *guarantee* to extract from each page. Of course, many listings have many more details listed than these which we can and plan to use.

As mentioned above, data will be scraped offline from saved .html pages with BeautifulSoup and will be structured into a pandas dataframe. Dummy variables for categorical variables may be generated and concatenated to this dataframe to facilitate use of these variables. Subsequent processing will be done using built-in pandas masking to get relevant rows from the dataframe for new queries when searching recently listed used cars.

Exploratory Analysis

We will visualize our data in multiple ways to check our data scraping procedures and make sure we did not incorrectly classify our data. The first basic check we will do is scrolling through the data frame for any obvious errors using the display command. We will then use the describe command to look at the descriptive statistics of each column in our dataframe. Next we will visualize the data using a scatterplot matrix in order to check the histograms of each parameter for outliers and general trends. We can also use the scatterplot matrix to explore correlations between different parameters. We will also visualize a heat map of the correlation matrix to determine which parameters are strongly correlated. This information will be used to identify potential strong predictors for the multiple linear regression and determine if any parameters are potential confounders.

Analysis Methodology

Regression
We will use regression to see if we can predict the price of a newly-listed used car. Our dependent variable will be list price and possible independent variables we will analyze include: year, seller type (dealer, private), mileage, color, title (clean/salvaged), transmission type, cylinders, fuel type, number of doors, exterior/interior

condition, listing date, page views per day (when a listing has reached 7 days). We will use the Python package statsmodels to perform all regression analyses. We will do a multiple linear regression first using the parameters that had strong correlations with list price. Based off of this initial model we will adjust our multiple linear regression to only include parameters that have significant p-values for their individual coefficients. We will use a significance level of α=0.05. Our final model should have a p-value < 0.05 for the F-statistic of the overall model. We are aiming to explain at least 70% of the variance with our model and hope to get an R-squared value of 0.70 or more.

Clustering
We plan to cluster what we classify as a “good deal” in its respective geographical location and create clusters showing areas in Utah where cars are generally sold for a good deal. As we have little experience here, we expect our experience to grow from the remaining material in the course.

Project Schedule

February 24th - 28th

- Check data accessibility (robots.txt and terms of conditions)
- Basic info due Wed Feb 26th
- Project Proposal due Fri Feb 28th

March 2nd - 6th

- Download html files for all recent listings from ksl
- Begin data scraping and create one dataframe with each row as a listing
- Get/give peer feedback March 5th
- Written feedback from staff by March 8th

March 9th - 13th (Spring Break)

- Finish data scraping
- Exploratory analysis
    - Describe

March 16th - 20th

- Exploratory analysis
    - Scatter Matrix
        - Interpret histograms - check if there are any outliers that could be an error from scraping
        - Interpret correlations
    - Heatmap of Correlation Matrix
        - Interpret Correlations

March 23rd - 27th

- Write up project milestone
- Project milestone due March 29th
    - Acquired, cleaned data, EDA, Sketches of your analysis methods, Submit zip file with Jupyter Notebook, data, other resources.

March 30th - April 3rd

- Get staff feedback
- Begin testbed for good deal predictions based on relation to scraped historical dataset

April 6th - April 10th

- Finalize predictive model for new listings
- Script and film project video

April 13th - 17th

- Polish up repository in preparation for final submission
- Edit and finalize project video
- Project Due Sunday April 19th
- Project awards April 21st