Presentation Notes

**Dataset**

This was an extremely large dataset containing Craigslist used vehicle listings from all around the US, there were nearly 500k rows in the initial dataset. We wanted to determine which variables in the listings were the strongest predictors of a vehicles price.

**Preprocessing**

The first thing we did was reduce the scope of our analysis by deciding to work with only used vehicles listings in Florida. The data contains 3 numerical variables (price, age, mileage) and many categorical variables. We removed variables that were not found to be significant for our purposes and then we had to do extensive cleaning of the data which contained many null values and values that made no sense. We simplified the car manufacturer variable to be binary (either domestic or import) Finally, the distributions of the quantitative variables were found to be extremely skewed, so we ended up taking the log of them which made them resemble normal distributions closely.

**Priors**

Many of the categorical variables either didn’t contribute much to price at all or didn’t provide much additional predicting power when including them. We found that condition and domestic were the strongest predictors among the categorical variables and we chose them, in addition to log(miles) and log(age) as out priors

**Histograms**

You can see here the effect of taking the log of price on its distribution, this holds true for the other quantitative variables too.

**Fitting the Model**

Here are our chosen priors in the model. The b1 – b2 coefficients correspond to log(age) and log(mileage) respectively. The rest are dummy variables that we generated for each factor of out categorical variables.

**Posterior Output**

This is our model summary showing each coefficient and the intercept, the next slides give a more visual depiction of what this means.

**Distribution of y-intercept**

This graph was constructed by sampling from the intercept and b1 coefficient of the posterior, you can see that it has a pretty nice distribution with no particular pattern or skew. We constructed similar distributions for the other variables, and these are what helped us decide whether to keep a variable in our analysis, many variables were extremely correlated with eachother and we decided not to include them due to redundancy

**Plotting Predicted vs Observed**

This graph plots the predicted log price vs observed log price and it probably the strongest indication of the strength of our model. You can see that past an observed log price of 8 the model is fairly symmetrical and seems to indicate that the model is reasonably strong. Unfortunately it does suffer below that number. There are many potential reasons for that given how noisy our dataset was. It is clear that despite our efforts, there were still quite a few outliers in the data for instance. Nonetheless, we can say that the model does in fact do its job of predicting price reasonably.