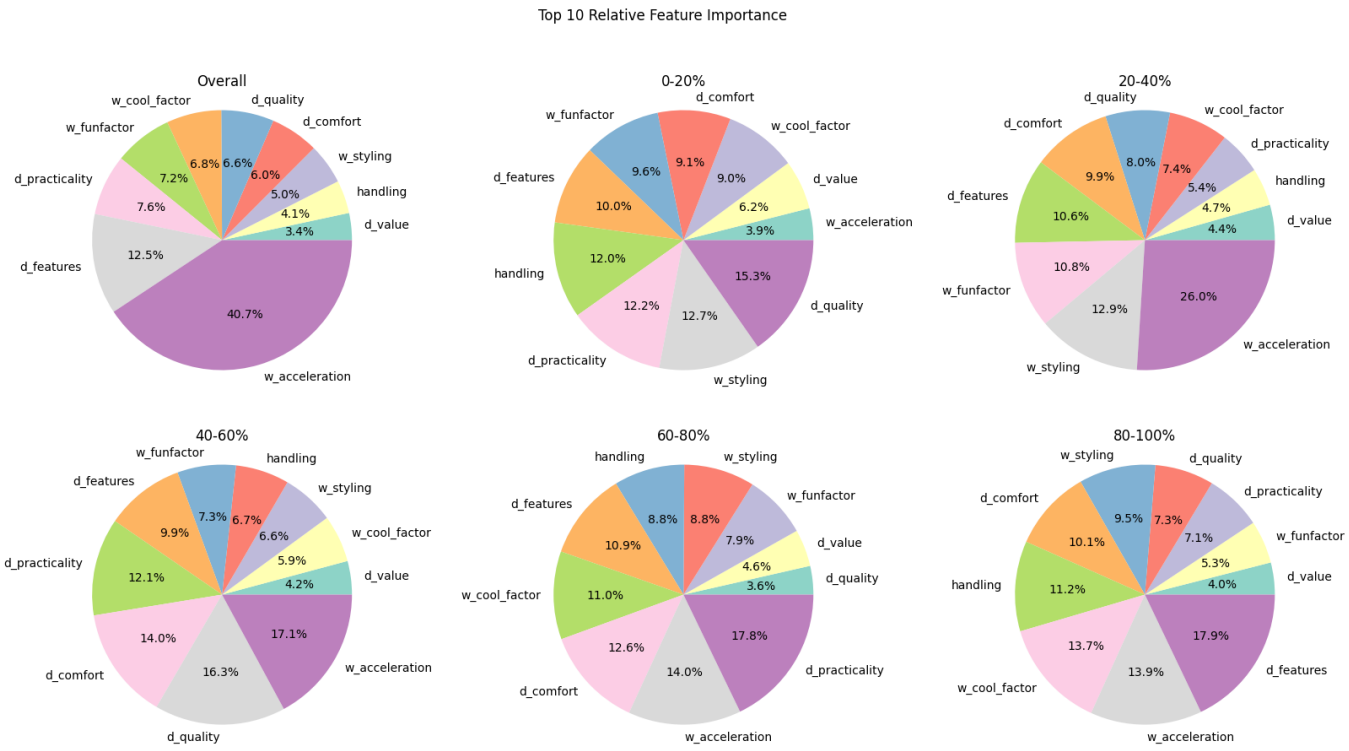
Doug Score - A look from an analytical angle

Introduction:

In this report, we'll dive deep into understanding the factors that influence Doug DeMuro's car reviews. Doug's reviews are quantified into the 'Doug Score', and our aim is to decode the elements that make a car stand out in Doug's eyes.

Key Findings Preview: Our analysis revealed that acceleration plays a dominant role in Doug's scoring, closely followed by car features. Surprisingly, aspects like car's age, or the filming location don't significantly impact the score relative to other features. There is impact, but explanatory power of the score than the other factors.

Doug's score really does vary in intensity when we look at the results of his high end and low end cars, what become the most important aspect shifts depending on what type of car we're looking at. Doug needs a way to adjust acceleration, its too prominent and he could benefit from scaling this feature so the importance is balanced across reviews. The same could be said for tiering the reviews between vintage and modern cars



Finally, there are a lot of features, but we found that you can actually make accurate predictions about what the Doug Score will be based off of as little as 2 - 3 factors, as we analyzed over 680 trees to find the features that were most predictive and were able to predict if a car was a

'Good' car or a 'Bad' car with over 90% accuracy and predict a score with only 7% error in the final value.

Where we're heading

In this report, I provide a quick insight into how Doug Demuro an avid car reviewer, might be evaluating cars. This analysis starts with a look at the review data, from the number of brands he's reviewed, the countries of the cars, the age of the car, as well as filming location, and duration. We'll take a look at all of the factors that go into the Doug Score based off of Doug's Score Sheet and our static copy used for this submission. These features listed provide the basis for his daily, weekend, and final Doug Score.

Once we've had a look at the data coming from the reviews and some of the light insight involved, we'll answer a few questions about the data that speak to the relationships between different features used in the review. Finally we'll use some machine learning techniques, to see if we can gain more insight into what Doug values in a car including sets of features that might see what makes a car a top Doug Score car.

Background on Doug Score

Here we'll go over the background on the Doug Score, what makes up a Doug Score, what do the attributes contribute, and we'll provide insights into those attributes

Scores at a glance

What's the score?

`Total Weekend Score` + `Total Daily Score` = `Dougscore`

Weekend Score

The weekend score consists of 5 features, that range from a score of 1-10

- Styling
  - How does the car look?
- Acceleration

Time Range	
Under 3 seconds	10
3.0 to 3.5 seconds	9
3.6 to 4 seconds	8

Time Range	Points
4.1 to 4.5 seconds	7
4.6 to 5 seconds	6
5.1 to 5.5 seconds	5
5.6 to 6 seconds	4
6.1 to 6.5 seconds	3
6.6 to 7 seconds	2
7.1 seconds and up	1

- Handling
  - the best cars will earn a 10, while the worst car will earn a 1. A “1” in handling will be reserved for cars that actually feel dangerous to drive
- Fun Factor
  - The ones with the highest fun factor score would be the ones I drive first, and the ones with the lowest fun factor score would be those I’d drive last
- Cool Factor
  - A combination of both how cool the car is and how important it is

Daily Score

The daily score consists of 5 features, that range from a score of 1-10

- Features
  - How well equipped it is. Only the cars with the most innovative features score very high in this category.
- Comfort
  - This category is about the smoothness, the ride quality, and the luxury
- Quality
  - Does everything feel nice? Has the car cut corners in any obvious areas? Are there any rattles or shakes where there shouldn’t be? Is the care reliable?
- Practical
  - Objectively measures storage space , but removes points or adds them for subjective things like ,how practical a car is to actually use and fuel economy

Cubic Feet Range	Points
0 to 3 cubic feet	1
3.1 to 6.5 cubic feet	2

Cubic Feet Range	Points
6.6 to 11 cubic feet	3
11.1 to 16 cubic feet	4
16.1 to 24 cubic feet	5
24.1 to 34 cubic feet	6
34.1 to 48 cubic feet	7
48.1 to 64 cubic feet	8
64.1 to 72 cubic feet	9
72.1 cubic feet and up	10

- Value
  - whether the car is worth its current market value

What other things do we know about the reviews ?

This are not quantitative factors, but we have some additional information about the reviews , we know the following

- Car Brand
- Model Year
- Country Filmed in
- Duration of the review
- City Filmed in
- Region Filmed in ex. State
- Country Vehicle was Made

How's the Doug Score calculated ?

We add up all the features from both categories and get the Doug Score

Doug provides two categories of score, a weekend score and a daily score. The combined scores lead to a Dougscore

`total weekend score = styling + acceleration + handling + fun factor + cool factor`

`total daily score = features + comfort + quality + practical + value`

`Doug Score = total daily score + total weekend score`

What do we really want to know

What influences the Doug Score and how!

Data Overview

Here in this section we will walk over what type of review data we have and some of the stats about the data that influence how we've gone about evaluating the Doug Score. If you want to skip the details feel free to look at the chart and keep scrolling!

Chart Breakdown

Attribution Score Distribution

We use a violin chart to show the how the scores are distributed for each attribute. The wider areas of the chart represent more values being represented in that range. The Y values represent the range of values for each score. The box with a line in the middle represents where you can see the average value of the score.

1. **Daily Attribute Score Distribution**
    - Practicality - has a few areas of more concentrated values in lower range
    - Comfort - there is overall more comfort in the cars he reviews
    - Value - is concentrated in a few ranges slightly above a 5
    - Quality - is concentrated in a few ranges slightly above a 5
    - Features - are slightly more concentrated in the mid range and almost distributed normally like a bell curve, with a skew towards large values
2. **Weekend Attribute Score Distribution**

- Acceleration - has a slightly more concentrated tilt towards being on the low end or high end
- Handling - skews slightly towards lower values
- Fun factor - has bell curve of distribution with more concentration at it's edges, meaning more values at the extremes of the score than with a typical normal distribution
- Styling - has a classic bell curve(normal distribution)
- Cool Factor - has a classic bell curve

3. **Dougs Attribution Score Distribution**

- Daily - has scores that average 28, with a top score of 40, and has another slight concentration around 34
- Weekend - has scores that average 28, with at top score of 49, and a more normal distribution with fatter tails, meaning more values at the extremes of the score than with a typical normal distribution
- Total - we can see its a sum of scores , where the average scores 57, and the max of 74, and are more concentrated above the average

Film Data

4. **Film Duration**

- Film duration is most concentrated in the 20-30 min range

5. **Film location**

- Most films were shot in the US, with a varying amounts in specific states, not shown here but weren't a contributing factor to scores according to models we tested.

Vehicle Data

6. **Vehicle Country**

- Germany - is an outlier Doug reviewed alot of cars from Germany, followed by the US, if we group them by large regions the major categories would be North America, Asia, Germany, Rest of Europe

7. **Model Year**

- Years - Doug reviews cars from 1950 to 2022, we bucketed them into decades, because it seemed logical and allows you to see he did the most reviews of cars less than 25 years old.

8. **Brands**

```
- BMW, Porsche, Ford, and Mercedes Benz are a bit over represented in the
reviews
```

9. **Brands another slice**

```
- We took the top 20 brands to look at according to review volume, no one
brand is represented more than 11% of the total number of cars reviewed.  So
it's a bit unbalanced.
```

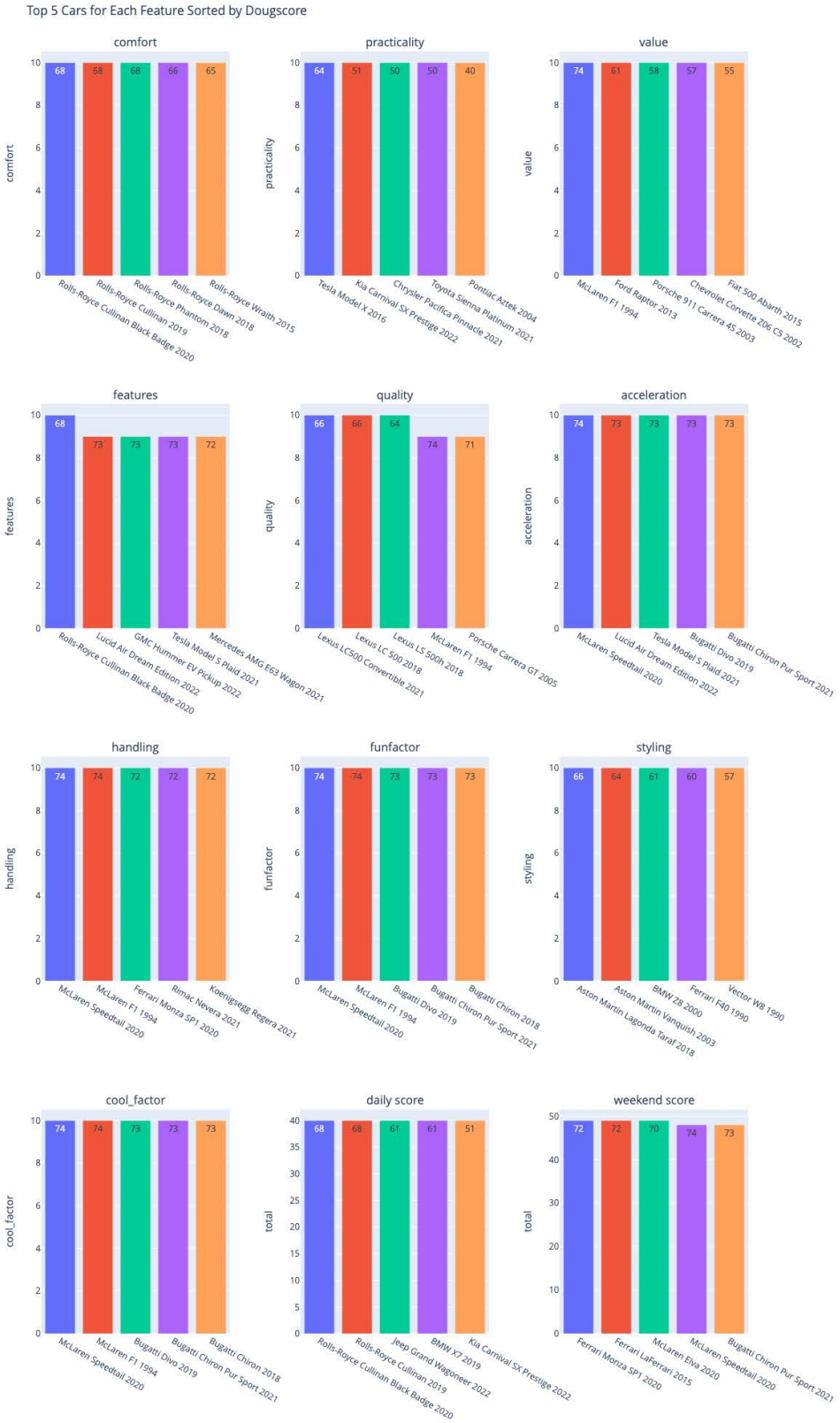
Relationship Analysis

Now that we have a good idea of what we're looking at when it comes to Doug Scores , let's take a look at the relationships between them specifically, we want to know what are the top 5 cars in each category. We also want to understand the relationships between the following attributes.

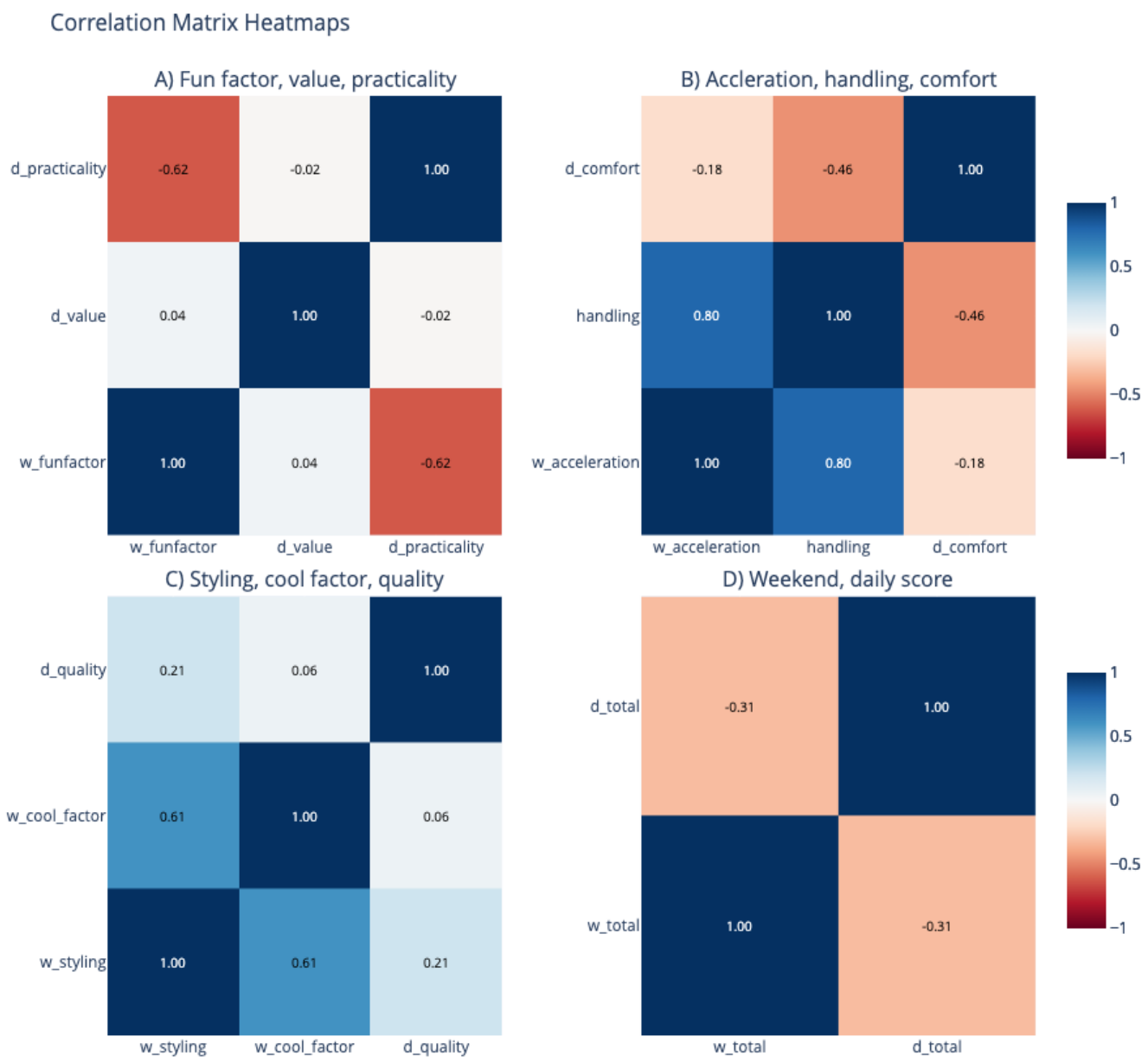
A) Fun factor, practicality, value
B) Acceleration, handling, comfort
C) Styling, cool factor, quality
D) Weekend score, daily score, brand
E) Weekend score, daily score, model year

In summary we took the approach of using pairwise correlation to explore the relationships between sets A), B) and C) and for D) and E) we used a regression model to see how much each factor was affected by their respective brand and model year. This allows us to also look at which elements of brand impact the score the most. The detailed results are in the relationship between features and regression analysis sections. In brief peak ahead, there are correlations between the feature sets and brand plays a significant role in the variability of the weekend and daily scores. Model year, however did not have much in the way of explanatory power.

Top five cars related to category



The Relationships Between Features



Here we take a look at what's driving the relationships, we choose to do a process known as pairwise correlation. Think of it of it as us saying what direction do the individual feature scores go in, do they go in the same direction + or in the opposite direction - , and by what extent. The values range from 1 being 100% correlated to -1 being 100% moving in the opposite direction.

A) Fun Factor, Value, Practicality

- We find that value and practicality only slightly positively correlated
- Practical cars don't seem to be a lot of fun

B) Acceleration, Handling Comfort

- Acceleration and Handling go hand in hand
- Being comfortable suprisingly goes in the opposite direction of handling
  C) Styling, Cool Factor, Quality
- it's interesting to see that quality isn't as positively correlated with coolness or styling
- Styling and cool factor make sense that they paddle in the same direction
  D) Weekend and Daily Score
- Weekend and Daily scores are negatively correlated which is a shocker that they move in opposite direction, but makes sense why no Dougscore in the 90s exist, as the max score you could have for any of the individual scores is 50 and if they're negatively correlated it means that both will never be in the same range at the same time. This makes Doug Scores
  hit lower

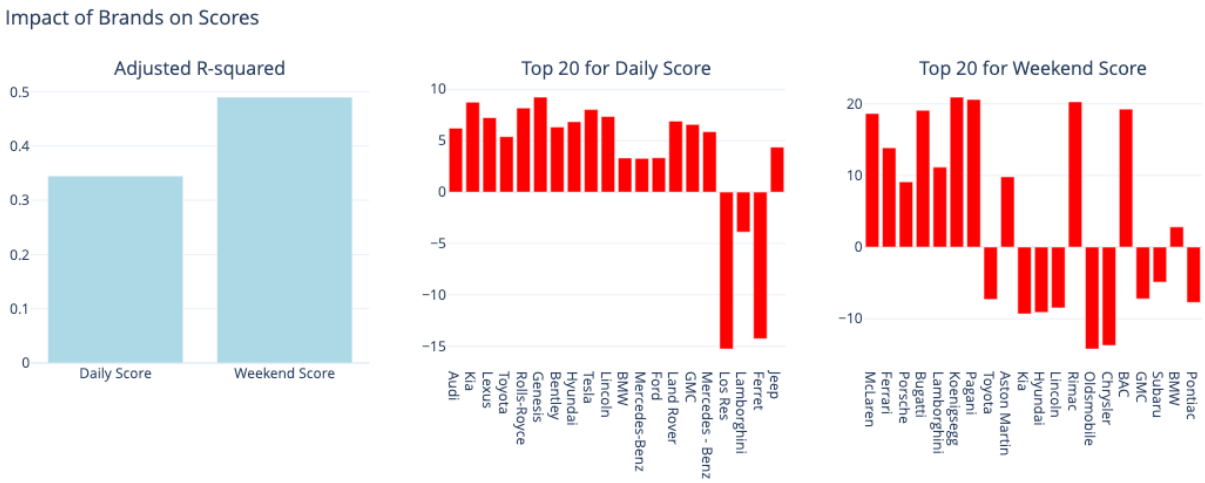
Regression Analysis for Brand and Model Impact

Here we took a different approach to measuring points **D** and **E** because we already know that the weekend score and daily score are negatively correlated moving in the opposite direction, we decided to do a regression analysis on brand and model year.

Adjusted R-squared is a metric we use to measure the explanatory power of our regression model, to capture the variability in the weekend and daily scores. Think of it as a number between 0 and 1, where the number closer to 1 means the more our model is able to explain the score.

In summary, we found that brand has a significant explanatory power related to the daily and weekend score , while model year across, different methodologies for grouping the years had little explanatory power on it's own.

Brand Impact



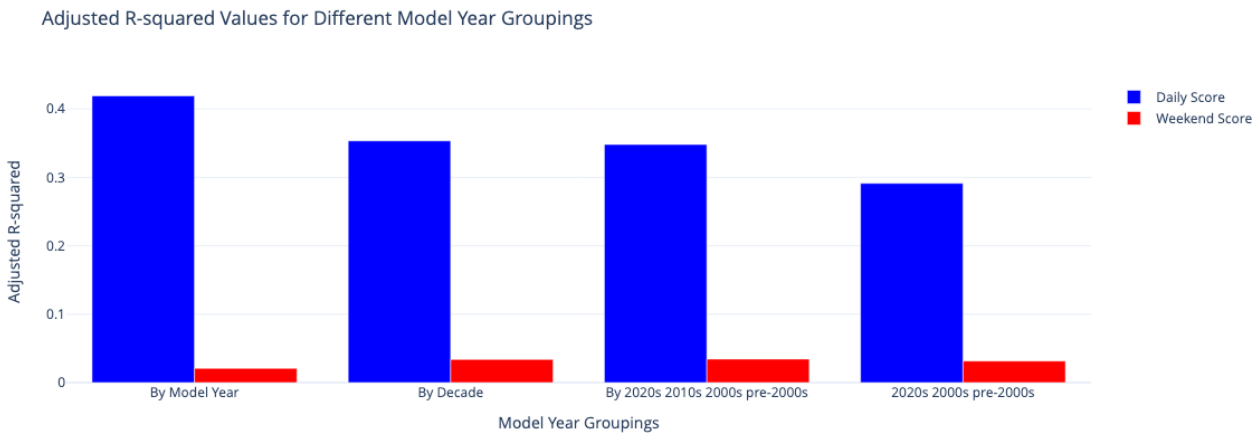
Here we find that with our regression on all the brands that we were able to explain the daily score variability with an adj. r-squared of 0.34 and the weekend with an adj. r-squared of 0.49. This implies that there is significance to brand in the score.

The Top 20 for the daily and weekend score show the coefficients (a measure of magnitude of impact) based off of the top ranking statistical significants of each brand item , for Example *Kia* contributes more to a high daily score prediction while , *Lamborghini* to a lower one

Model Year

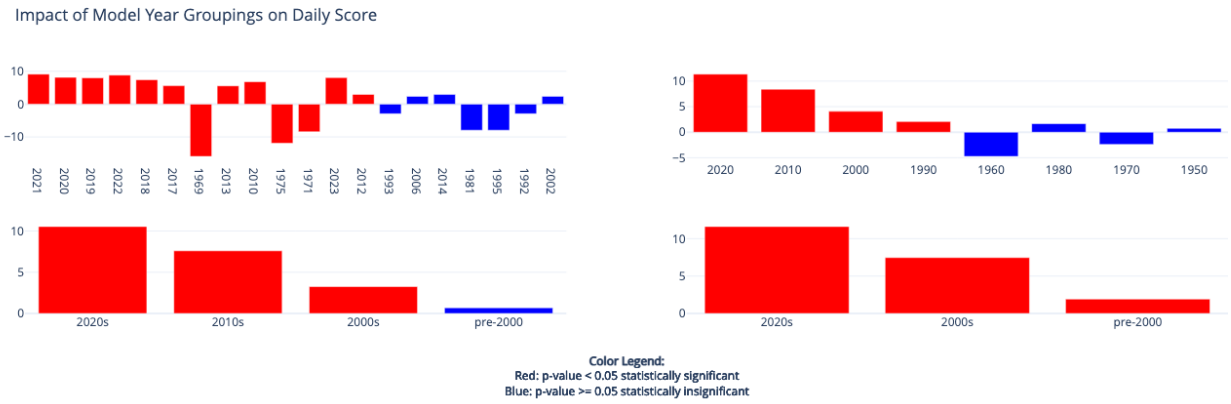
With the model year it being a large space, we took several approaches to the data, when dealing with regression models, the more variables the more the variability is explained, so when thinking about scoring we thought it'd be good to look at the model year in buckets as well as absolutely to gauge it's impact.
We created 4 groupings,

1. By Model Year - strictly the year
2. By Model Decade - 90s, 80s, 50s etc...
3. By 00s,10s,20s, pre 2000s - This was done to balance the cars in each category
4. 20s, 2000s, pre 2000s - This was done to show a grouping between contemporary cars, modern cars, and super vintage cars



Here the adjusted r-squared by every grouping for model year has some explanatory value to the score variability in the daily score and very little for the weekend score. For this we kind of discount the specific model year as this will probably be too specific to say oh 1975 means a good car year. So we extrapolate this to different logical pairings and we still see some explanatory power for the variable.

Impact of Model Year



The Top 20 year entries for the daily score show the coefficients (a measure of magnitude of impact) based off of the top ranking statistical significants of each model year bucket , for Example *2020s* contributes more to a high daily score prediction while , *pre-2000* is statistically insignificant in a 4 category model year grouping.

Modeling the Doug Score

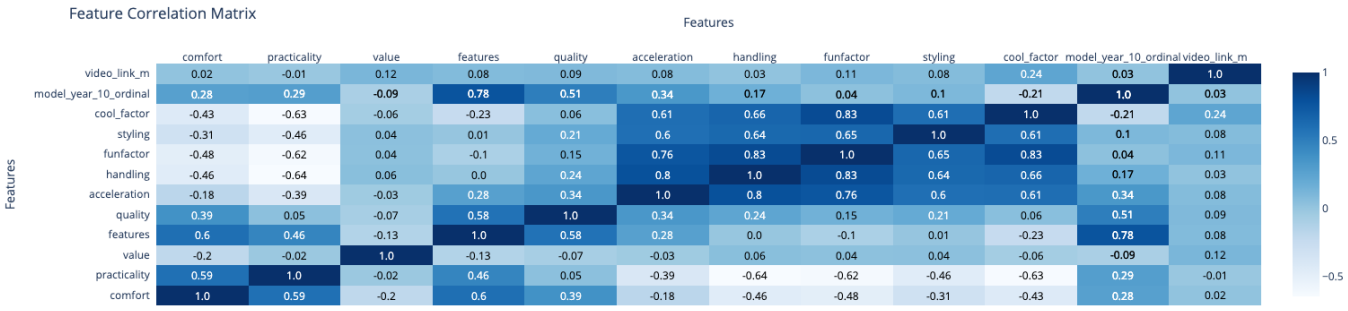
Deciphering the Doug Score is akin to unlocking the mysteries of a complex puzzle. Each component of the score plays a unique role, and understanding their interplay is essential for grasping the bigger picture. What drives the overall score? How do individual elements converge to shape it? Let's embark on an analytical exploration.

Our primary aim is to dissect the intricate relationships between the various scoring components. We want to ascertain how they collectively mold the Doug Score and discern the underlying rules governing it.

To achieve this, we employed a mix of analytical techniques, primarily focusing on decision trees. These trees, which range from straightforward binary classification models to more advanced gradient-boosted regression trees, became our primary tools for investigation. The results were illuminating. Even without leveraging all the features, we managed to predict Doug Scores with over 90% accuracy and a mean absolute error of just +/- 7%. To put this into perspective, even a sophisticated model with 5000 iterations and all 10 feature sets only reduced the error to 5%.

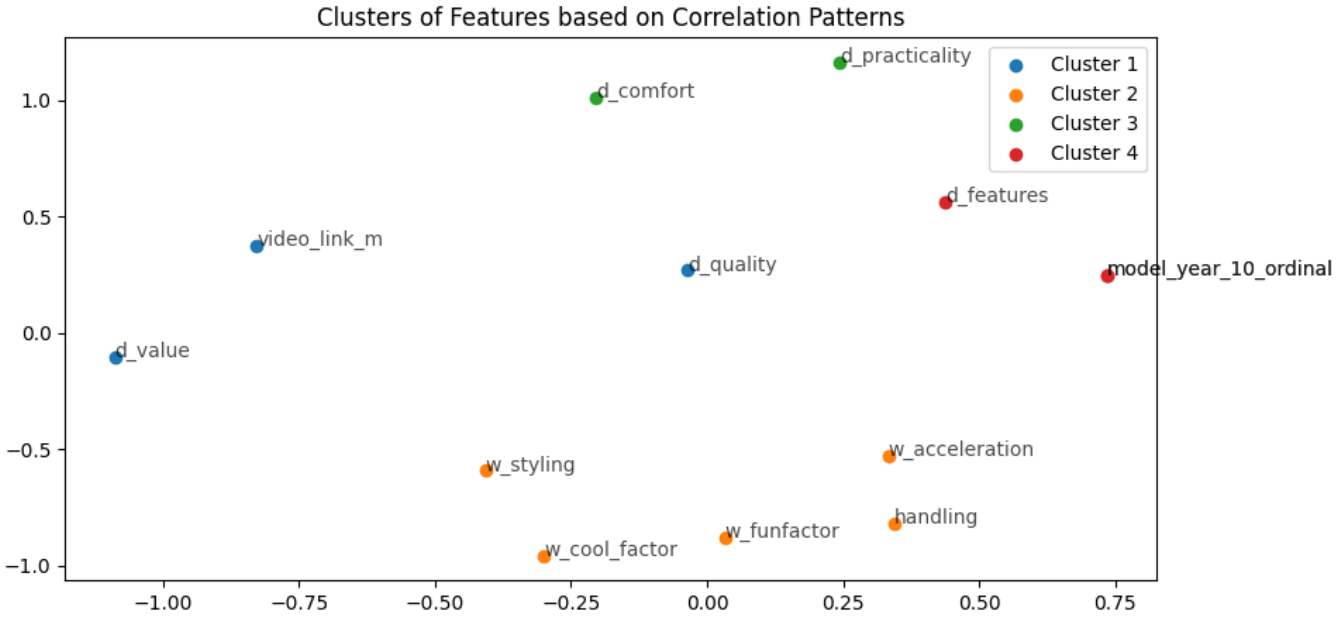
What does this tell us? The essence of the Doug Score, it seems, can be distilled into a few key factors. Notably, the components `acceleration`, `features`, `funfactor`, and `quality` emerged as pivotal influencers as well as the `model year decade`. Join us as we delve deeper into our optimal tree model, shedding light on its intricacies and interpretations.

Feature Data



Here we've created a heat map to show how all the features are related to each other(correlated), this will help us to understand how we can use it to understand what features might have interesting relationships. The set of features that will be most expressive together will be less related to each other. That's what we're looking for

Where to start?



We wanted to know how these correlations were related so we performed a process known as KMeans clustering on the correlations to figure out what features of Doug's review are related to each other the most. We used a analysis technique called `elbowing` to figure out the number of clusters that would best capture our features.

Cluster 1 - Video link(duration) , `value` , and `quality` all have some factors that are related , and `features` and `quality` are some what peddling in the same direction, as also indicated by the heatmap

Cluster 2 - `Acceleration` is at the center and `handling` , `funfactor` and `coolfactor` are more closely related to each other with styling being correlated but a bit further away from `funfactor`

Cluster 3 - `Comfort` , `practicality` grouped together

Cluster 4 - `Features, model year decade (model_year_10_ordinal)` makes sense that the two would be related as the larger the decade number the more recent the car so the larger the number of features

This helps guide our intuition as to what should make an expressive model.

Modeling Insights

Simple models are the best models, they're the most expressive and the most generalizable. When we add features we often run the risk of missing the importance of factors. We found that using only 2 or 3 features of the 15 features available including groupings of model years, brands , varies configurations of vehicle country origin, Just a small handful performed as well as an almost full feature model, resulting in 1% drop in predictive quality.

Out of the 680 distinct decision trees we generated over pairs and triples of features, we found that these triplets performed the best

Classification (Good or Bad Car) > 90% accuracy:
three-feature:

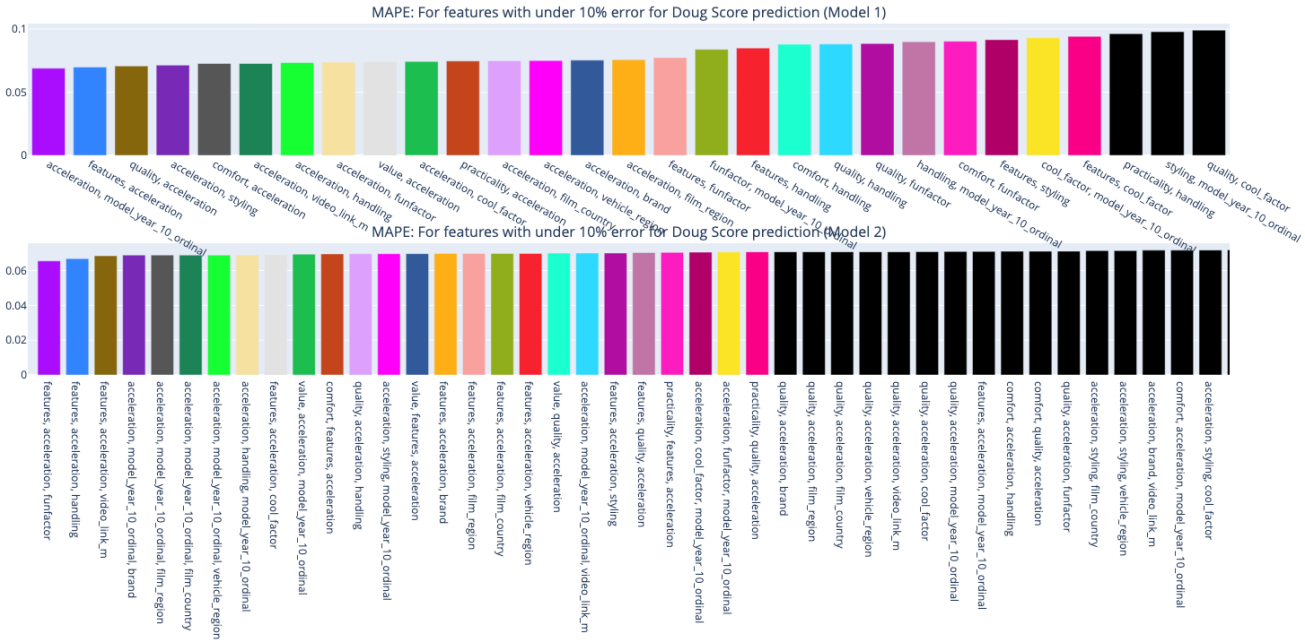
- features, acceleration, handling
- features, quality, acceleration
- comfort, features, acceleration
  two-feature:
- features, acceleration

Regression (Score Prediction) < 7% error:
three-feature:

- features, acceleration, funfactor
- features, acceleration, handling
- features, acceleration, video_link_m
  two-feature:
- features, acceleration
- quality, acceleration

Model Performance

Regression Decision Tree Model Performance

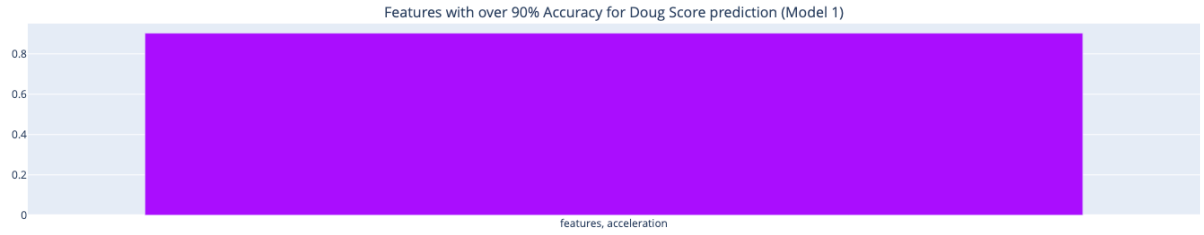


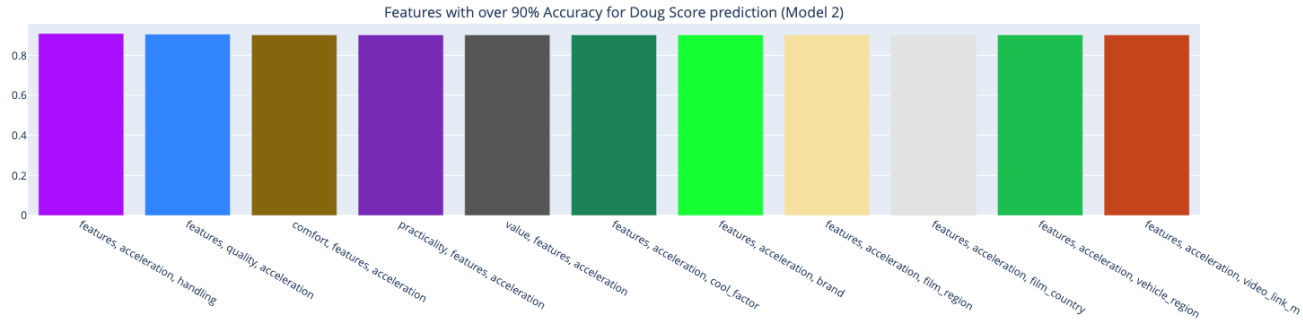
This shows our binary decision tree performance for 'good' or 'bad' score depending on the Doug Score range. The top chart is our top performing two feature model, and the bottom or top performing three feature models. This measures the prediction accuracy of classifying a car as good or bad depending on if it is in the top 50% of Doug Scores based on 2 or 3 features. We used shallow trees, with a depth of 3 levels, meaning the rules are less fitted to the data and more general.

Our top 3 feature model for binary classifier was `features`, `acceleration`, and `handling`

Decision Tree at a Glance

A Quality, Features, Acceleration Regression Decision Tree

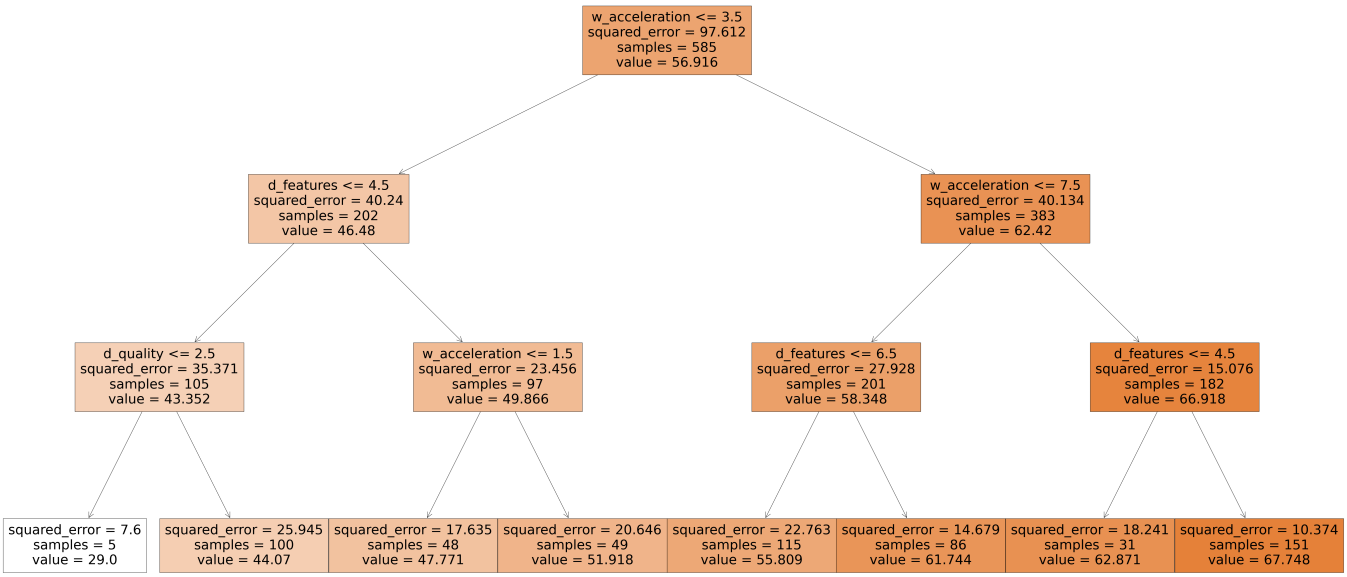




This chart shows the top performing regression decision tree models trained with those feature sets and their error rate. The top chart shows the two feature models, and the bottom the three feature models. It shows less than 7% error rate. This means that using two or three values from the review it can predict the Doug Score within 7% most times.

We used shallow trees, with a depth of 3 levels, meaning the rules are less fitted to the data and more general. The top 3 feature model was one of `features`, `acceleartion`, and `funfactor`

Classifier Decision Tree Model Performance



This simplified decision tree achieves 7% error rate, it allows us to see the how Doug could be making his pricing predictions more generally, in terms of importance. This shows that we really could predict the Doug Score with some accuracy with just 3 features and a short tree.

How it works

Starting at the top of the tree we have the feature that is most informative of what the Doug Score will be. When accceleration is less than 3.5 or greater than, the model is telling us that we gain the most information about what the Doug Score will be . If that's the case we will find ourselves with a score that is less than 51 the lower half of the Doug Scores otherwise we'll find ourselves closer to the upper half. We simply follow the paths until re reach the last node. The last node is the prediction of what the price will be. The more depth, the more granularity we get in the final score value, and the more nuanced the decisions get.

What does it mean

We ran over 680 trees, with a very shallow depth of 3 levels, and were able to achieve a 7% error rate in our predictions. Decision Trees enable us to quickly assess the importance of features to the Doug Score. We found that acceleration across the entire dataset is the most key factor followed by features. Because we generated every combination and were able to measure the performance of all the trees, we were able to understand deeply what makes the Doug Score tick.

How do we confirm

Our next approach was to really dive into the trees to see what we could find, we used a technique called XG boost which essentially generates 5000 trees, to figure out the best decisions on the data. We also did this on the full data set. We did this to see what features,

the model would thing were the most important. We then computed the relative importance of each feature in a pie chart.

Why

We did this to make sure that we could then with full data understand the interactions between all the features including the categorical ones that made sense for a given tree depth. Because there are 1000s of trees that are combined it's hard to visualize , so representing an expressive smaller tree, with comparable performance gives us a good insight into what's happening.

Questions & Results

At the beginning we had a few questions about the interaction of data and Doug's score. With the analysis and tools we've put together we can now answer a few questions about the data.

Do certain components of the DougScore disproportionately influence the final score when interacting with other components?

Yes, we find using our decision tree analysis with our complex model, across all the data that `acceleration`, `features`, `practicality`, and `funfactor` were the most impactful to the model. This is evidenced by the following relative importance pie charts as well as our smaller 3 feature models, whose prediction error was minimized by simply choosing some combination of those factors.

Is the contribution of individual components to the final DougScore consistent, or does it vary (perhaps diminishing or plateauing) at different scoring levels?

Yes, actually and it intuitively makes sense, for scores in the lower quintile of the Doug Score, we find that those cars are all slow, so acceleration isn't a meaningful attribute into figuring out what the Doug Score is the features shifts to `quality` and `styling` , vs `features` which for cars at the top is the primary determinant of a great score, as most cars are already fast and high `quality`

Investigate whether and how different scoring components interact to impact the final DougScore. For example, does a high acceleration score contribute more to the overall score when accompanied by a high handling score?

We took the approach of performing a regression analysis with interaction effects on feature pairs of size 2. We found that there are statistically significant meaning that the interaction of the pair meaningfully impacts the resulting Dougscore when used as dual factors in a model.

We also found that the affects are non-linear, we binned the Dougscores into quintiles and found that for instance `acceleration` and `handling` have positive interaction across the

data set but for the 80-100%, but specific quintiles like the 0-20% have a negative interaction effect, meaning that the affect is less than what you'd expect from the individual effects of the components. The size of the coefficient value could suggests the impact is small depending on it's relative magnitude to the rest of the components . We can say there is definitely a relationship that shifts over the quintiles.

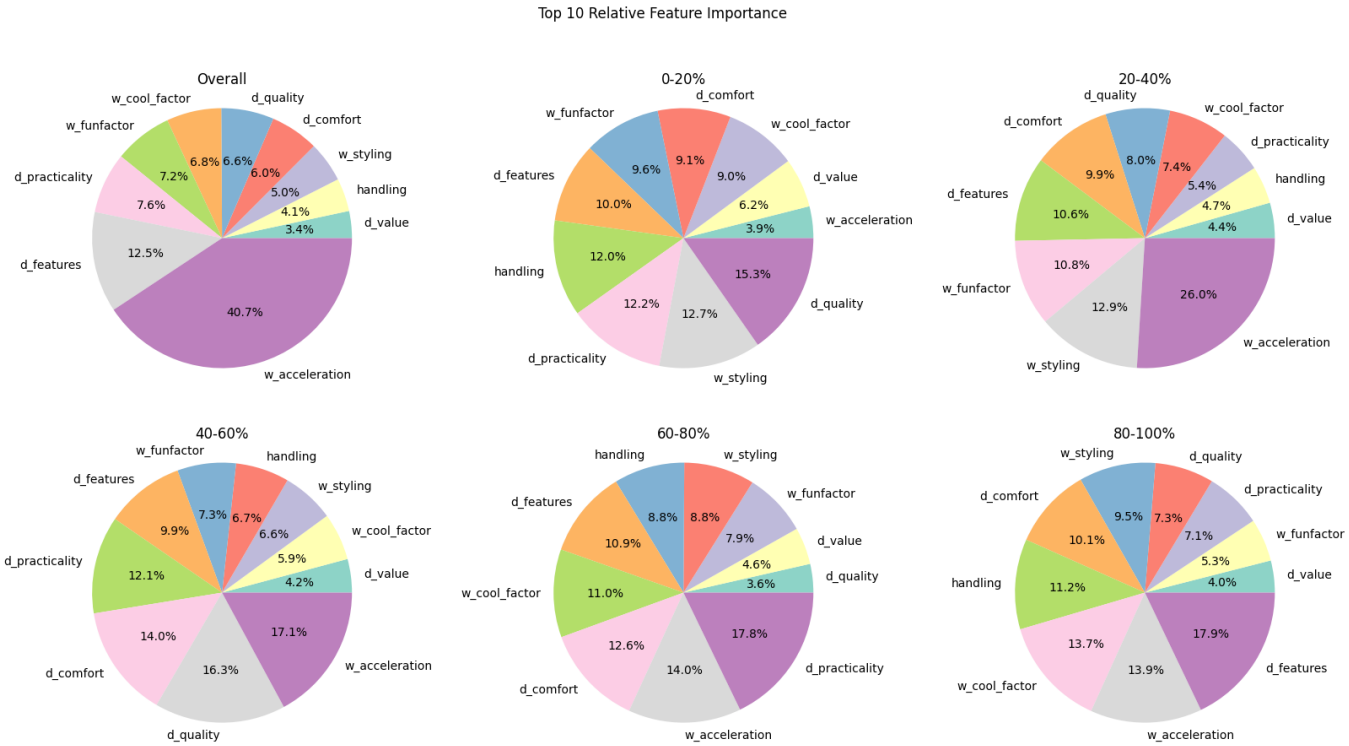
Significant Interaction Effects for 80% - 100% Percentile Rank

Feature 1, Feature 2 - 80% - 100% Percentile Rank	coefficient-value
d_comfort, d_quality	-0.06390533208899028
d_practicality, d_features	0.07137551110356201
d_value, d_quality	0.07664356747007874
d_value, w_acceleration	0.172367327586162
d_value, model_year_10_ordinal	-0.06047456830295747
d_features, d_quality	-0.12994679743642013
d_features, handling	-0.10607567530192336
d_features, w_funfactor	-0.0914431179298083
d_quality, handling	0.07459905311196031
d_quality, w_styling	0.09028431505294487
d_quality, w_cool_factor	0.10698144136232861
d_quality, model_year_10_ordinal	-0.18597431403668413
w_acceleration, handling	0.15600619280714048
w_acceleration, w_funfactor	0.18495054551017306
handling, w_funfactor	0.09768072704522349
handling, w_styling	0.08019241644453452
handling, model_year_10_ordinal	-0.10549472134306637

Examine the relationships between individual scoring components and the overall DougScore for potential nonlinear patterns. For instance, does the influence of acceleration on the overall score plateau or diminish after reaching a certain threshold?

There are tiers of car performance according to Doug that have different effects, when we use our regression trees, we can take the Shapely values to see that the affect on the components in terms of relative importance shift dramatically with each quantile,

highlighting the importance of specific features to specific ranges of Dougscores.

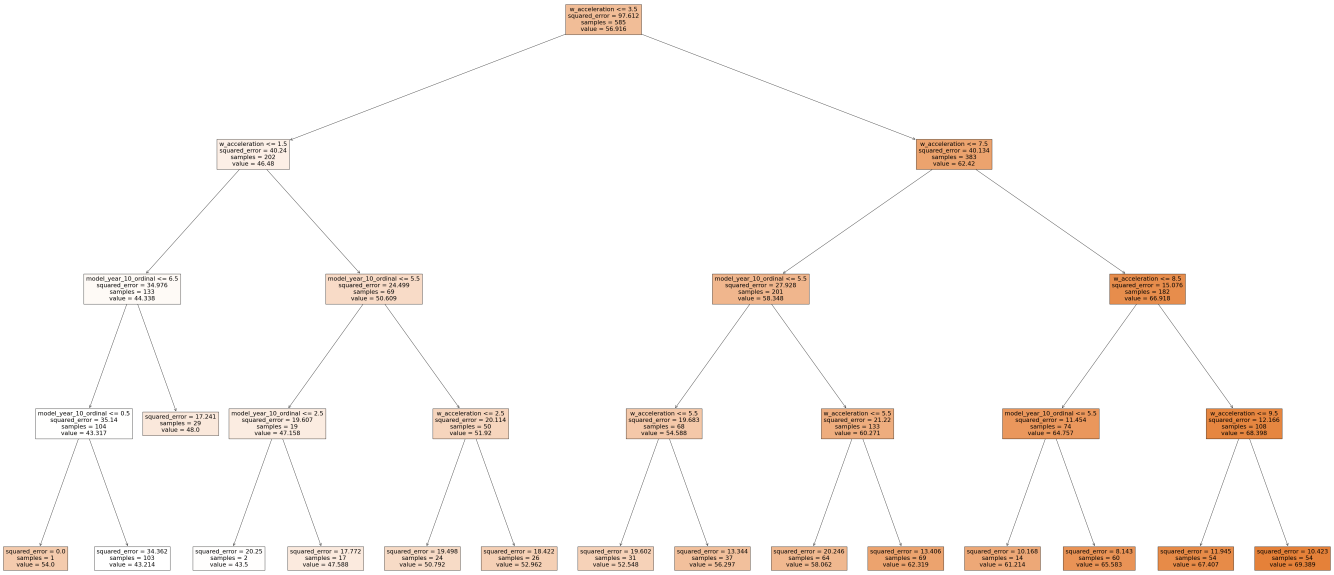


Here we break things down by quantile, and relative feature importance using a metric known as a shapely value, we take this value which measures feature importance, then we calculate how much it's shapely value is relative to the shapely values across the entire dataset, and that gives us the relative value for the plot.

This tells us that the valuation of each feature is not linear across the entire, range of Doug Scores. This means that for instance across the data set, the relative importance of `acceleration` is 40% compared to features, but when we look at specific ranges, like the high end portion of the Dougscore we see that features have an relative importance of 17% compared to accelerations 13%, implying that there is a different relationship for different segments. On the low end for instance acceleration accounts for 3.6% importance

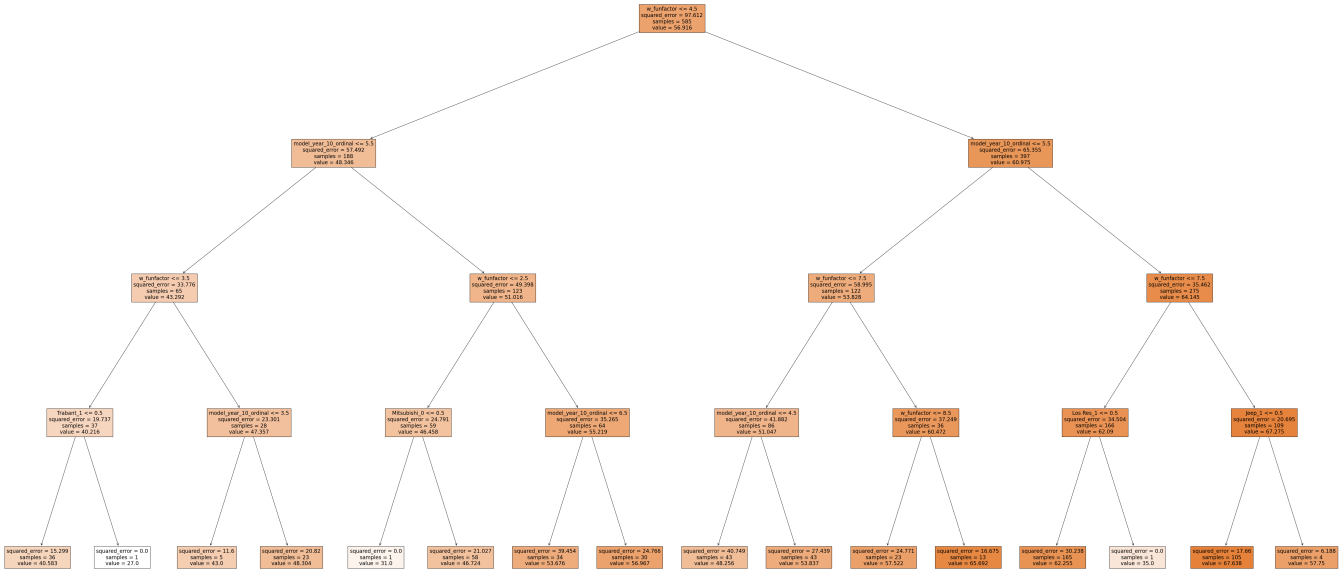
Additional relationships

Model Year By Decade And Acceleration - The fact that we can build a model ,
that is fairly predictive within 10% , MAPE, with these two objective features,
implies that the Doug Score could be heavily influenced by or explained by the year the decade the car was made and the acceleration it has.



From here we see that using this 2 factor tree of depth 4, we see Doug has a little more compassion for Modern cars with low acceleration, as in the lowest category, modern cars out score faster less modern cars. It could be because of features, but it's interesting to observe that they consistently score lower, even in the face of Doug's cardinal sin according to his ranking order slow cars.

I wanted to see the relationship between Fun Factor and Model Year, does Doug have a penchant for vintage cars that are fun?



We performed a decision tree just on the features fun_factor and model_year, with a depth of 4, with a MAPE or error of 8% . We found Doug still loves his modern car more than his vintage ,

given similar levels of fun. However, when we look at cars that are just ok fun , a vintage car from the 80s-90s(3.5-4.5) is just a little better than a modern car that's completely no fun.

Conclusion

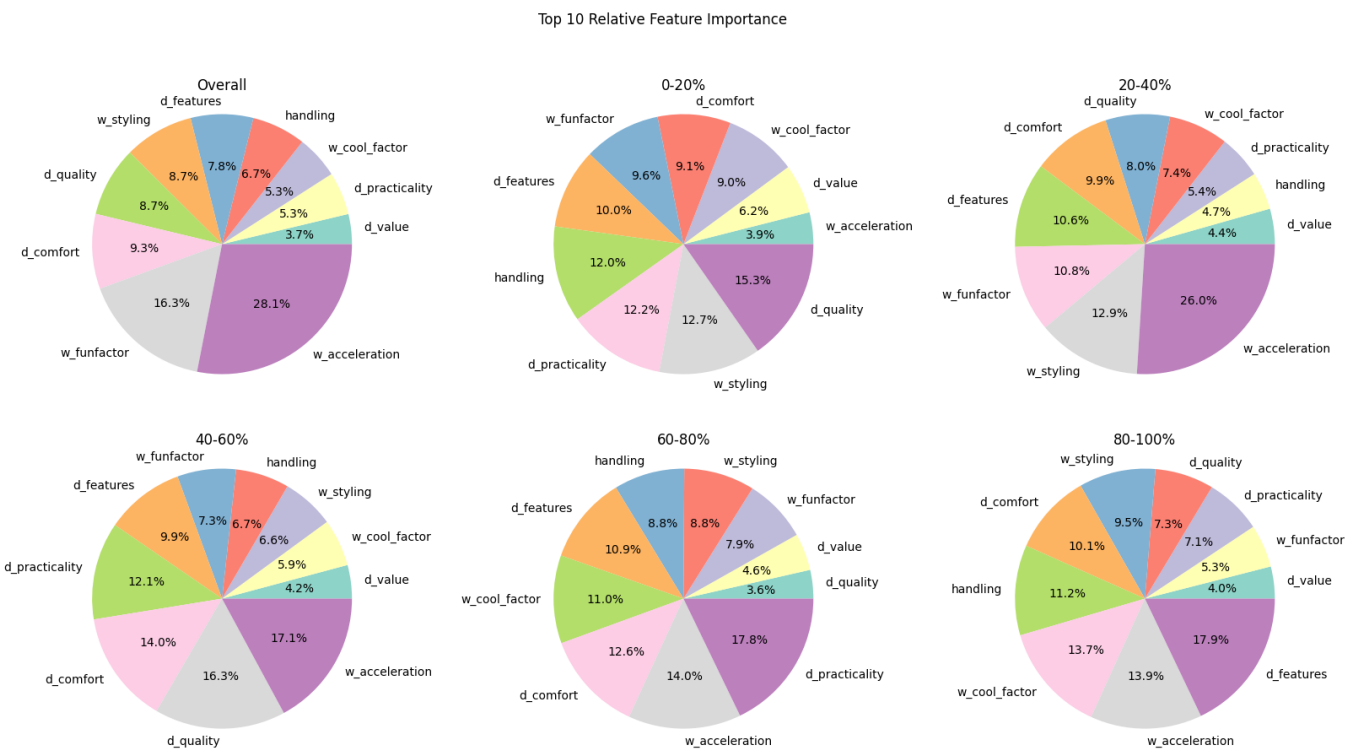
The Doug Score, is a really cool score, with lots of facets to it. Doug's score actually puts too much weight on things like acceleration, and if he penalized modern cars , or just judge modern cars, we'd see a different set of scores. Overall though, Doug's done a great job of getting a pretty diverse set of cars to review.

What's really interesting is that acceleration and features are such strong indicators, that modern cars are definitely going to pull the Doug Score upward, just because cars get faster over time not slower.

Doug should tier the cars into oldies (pre-2000), vintage (past 25 years), and modern cars
For vintage cars alone across all quantiles, the data is much more balanced for feature importance

That said really acceleration needs balancing as a score, it weighs to heavy , if we cut the feature points by half we'd maybe be in a better way.

Vintage car feature importances



Appendix:

Methodology

Preprocessing Data

Data loading

We first downloaded the csv data set, and then went about transforming each of the named columns, into something more machine readable, prefixing all but one of the features with a w *prefix for weekend attribute, and a d* prefix for daily attribute, with handling being a daily feature without the prefix.

The resulting headers are as follows:

model_year	brand	model_name	w_styling	w_acceleration	handling	w_funfactor	w

Data transforming

The key transformations that we applied were to try transforming all categorical entries into one-hot encoded values so we could use them across various models. transforming `model_year`, `vehicle_country`, `brands`, `film_country`, and `film_region`.

We synthesized the following data to, we transformed video_link a specific minutes:seconds duration to 10 minute intervals, to make the data less specific and more general, so we could use it in our models.

We also synthesized 5 different model year categorizations. Binning the model years into logical groupings such as model a 3 time period model, 4 time period model,
and a decade model. We also transformed the data, from a continuous specific year to a piecewise decade model, which would allow us to use ordinal data for the model year in our decision trees. This would also put the year data on the same scale as the rest of our features, making the data fit better in our tree.

For Vehicle country field, we found that country might be too specific as there were several countries that had a larger representation in the data set. In order to balance this feature we created a new category called `vehicle_region` that bins the vehicle country into `Germany`, `North-America`, `Asia`, `Other Europe` . This in turn turned the distribution of cars per country into an almost even distribution between the groups, making our model potentially more general.

All categorical fields were one hot encoded.

Missing Data

For the missing data, with the video link time we took the approach of just back filling it with the median time for the video duration, this allows us to not really shift the distribution of minutes much.

Country, Region

We did with regards to film and film location transform some of the regional data as duplicating that with say UAE vs the specific city something was filmed in as states and regions don't translate across countries. This didn't seem to have alot of importance to the review.

Data Analysis

Data Overview

In the data overview section, we wanted to see the distribution of the values , while also not ignoring the magnitude or the mean. We thought a violin plot would capture the data the best, wit this we're able to look at how the distribution of the features and their scores might be skewed or look, including if the distribution were normal or not, which it looks multimodal

The rest of the statistics were calculated as straight distributions and bar plots, we did transform Bran into a pie chart distribution of the top 20, there were 84 brands and the visualization would have been unweildly so we chose to jsut show the smaller portion.

Correlation Stats

For this section of the paper we computed correlation matrices for each of the sets of features. For the Weekend and Daily score, and the categorical features, we opted to do pairwise correlation on only the weekend and daily score, so we could ascertain if the two were correlated, which they were.

Interaction Effects

To measure the interaction effects, we computed an ordinary linear regression across the pairs `weekend_score` and `brand` as a one hot encoded feature and `daily_score` and `brand` as a one hot encoded feature. We displayed the R-adjusted score and the coefficients sowe could gain a look into the magnitude/direction of the impact of the feature values . R-adjusted was definitely necessary for brand given how many brand variables their were. We filtered the coffiecient results for statsitically significant variables with a $p < 0.05$

To measure model year and `daily_score` and `weekend_score` , we took 4 different approaches,
we couldn't be sure if binning would have an affect, and we knew that just picking years would

potentially lead us to over fit to the 1 instance of a 1991 car , so we binned and computed the r square adjusted for each , as well as plotted the coefficients for significant variables

Modeling w/ Trees

In overview we created an approach to understanding the Doug Score with Decision trees. We choose this because given the nature of the problem, we wanted to be able to visualize the learning model, and decision trees seemed like a good fit to be able to capture the linear and non linear relationships between variables in an interpretable way.

We created two types of trees, across 3 approaches. Starting from least complex to most complex.

Binary Decision Tree

So first we wanted to know if Doug's score was learnable so we decided to just see if we could accurately beat 50% with a `Good` car `Bad` car classification system. We created a 2 class output for the Dougscore, the top 50% were good cars the bottom 50% were bad cars.

So the first thing we did was to use scikit `sklearn's` decision tree classes
```

```
clf = DecisionTreeClassifier(max_depth=tree_depth, criterion='gini',
random_state=42, min_samples_leaf=10, min_samples_split=10)
```

Here we set the criteria to be gini so we can get a measure of how much each node was contributing, and then we didn't want to over fit so we set some reasonable sample size splits for the leaf and split nodes.

We were able to achieve 91% accuracy and we did this using stratified kfold cross validation of size 5, and with a tree depth of 3. The optimal we found to be actually deeper around 5 levels, but 3 was already expressive enough we felt to favor easy readability and interpretablity

Regression Decision Tree

For the regression analysis, we felt that since it's a continuous variable we can try and predict closely to that. We found a similar tree depth of 5 would be optimal but also found not much increase in error between a depth of 5 and 3 within from 5%-7%.

```
  # Train the regression tree
    reg = DecisionTreeRegressor(max_depth=tree_depth, random_state=42,
min_samples_split=10)
```
```

```
    kfold = KFold(n_splits=5, shuffle=True, random_state=42)
    cv_results = cross_validate(reg, X, y, cv=kfold, scoring=
['neg_mean_squared_error'], return_train_score=True)
        # Fit the model and make predictions
    y_pred = cross_val_predict(reg, X, y, cv=kfold)
```

So here we actually use min samples split to 10 and then our depth is 3 and then the scoring for cross validation was MSE , we then also wanted to know the MAPE so we could have a number that was easily interpretable for the consuming public.

Regression Sweeps

So after playing around with the features, being fully specified to understand the dynamics of the tree , we then asked the question what if we just found if we can create great predictions with fewer data points. We then permuted through all subsets of 2 and 3 features. We then some code to run our training data

```
#train and plot subsets
for subset in two_feature_subsets:
    two_feature_results.append(train_and_plot_tree(df,
daily_columns=subset[0], weekend_columns=subset[1], meta_columns=subset[2],
criteria=None, quantile=2, tree_depth=3, show_plot=False))
    two_feature_regression_results.append(train_and_plot_tree_regression(df,
daily_columns=subset[0], weekend_columns=subset[1], meta_columns=subset[2],
criteria=None, tree_depth=3, show_plot=False))

for subset in three_feature_subsets:
    three_feature_results.append(train_and_plot_tree(df,
daily_columns=subset[0], weekend_columns=subset[1], meta_columns=subset[2],
criteria=None, quantile=2, tree_depth=3, show_plot=False))

three_feature_regression_results.append(train_and_plot_tree_regression(df,
daily_columns=subset[0], weekend_columns=subset[1], meta_columns=subset[2],
criteria=None, tree_depth=3, show_plot=False))

classifier_prefix = "Features with over 90% Accuracy for Doug Score
prediction"
regression_prefix = "MAPE: For features with under 10% error for Doug Score
prediction"
```

With this we iterated over all possible subsets and only returned sets that were able to pass 90% accuracy or under 10% MAPE, which is where the charts come from. We then sorted them to display what was the most relevant combinations.

Other Techniques

We wanted to go for interpetablility and high accuracy, so we played around with a random forrest, but were only able to achieve very modest increases in performance per 100s of trees per set of features. So we simply moved on , we did the same thing for regression and classifier trees.

More than 2 buckets

We also swept across multiple combinations of quintiles for the classifiers, but with us bucketing and the decision tree, we simply weren't able to achieve the results that would have made a great classifier , that being greater than 90% accuracy on the validation sets.

Gradient Boost Trees

So finally we decided to use `from sklearn.ensemble import GradientBoostingRegressor` gradient boost tree, to model our data. We simply added all the features in, because we knew we weren't going to be able to visualize the trees, and opted for minimizing overfitting, while at the same time being able to understand how the model prioritizes features. So we used this to figure out feature importance per quintile of Doug Score data. So that means we looked at the ranges from the 0-20%, 20%-40%, 40-60% ,60-80% and 80-100% percentile

```
gbr = GradientBoostingRegressor(n_estimators=n_estimators,
min_samples_split=20, learning_rate=learning_rate, max_depth=tree_depth,
random_state=42)
cv_results = cross_validate(gbr, X, y, cv=5,
scoring='neg_mean_squared_error', return_train_score=True)

y_pred = cross_val_predict(gbr, X, y, cv=5)
mape = mean_absolute_percentage_error(y, y_pred)
```

We choose 5000 estimators, split of 20 , a learning rate of 0.01 (we tried smaller and larger, but didn't do a grid search, and a tree depth of 4. Again we choose cross validation for 5 sets, and to compute the MAPE for ease of use.

We also performed shapely analysis, here we use the shapely score, to be able to ascertain importance in the tree. We took this a step further by creating a relative shapely score metric, that would give us some idea of the importance.

```
relative_importance = mean_shap_values / np.sum(mean_shap_values)
```

This allows us to plot a pie chart of the top 10 features amongst all of them so you can get a since of the relative scale of impact.

We also toke the values from the regression and using the `shap` package plot the shapely values for each quintile

Clustering

We also took a stab at figuring out the relationships between the features using clustering. We performed K-means clustering on each of the features correlation, after transforming the correlation matrix to a strictly greater than 0 distance matrix. We then used `from sklearn.cluster import KMeans` and `from sklearn.manifold import MDS`
To compute the clusters, and then used MDS to reduce the dimensionality so we could plot them and see the groups that arose

Interaction Effects

In the interaction effects section, we took the approach of using a linear regression with `ols` `import statsmodels.api as sm` from `statsmodels` . We computed the following interaction effect
`output ~ feature1 + feature2 + feature1 * feature2`

The challenge here was to normalize the data and the features, by multiplying the features we change the scale, so we normalized the data using `StandardScaler` to then bring the interaction term to be at the same scale. We then went and plotted the tables for the interactions between
all pairs of `weekend` features as well as `daily` features with an additional term of `model_year_10_ordinal` our transformation of the model year to a scaled number corresponding to the decade since the earliest car review ranging from 0-7.

This allowed us to compute a table that lists out all the significant interaction terms with p-values < 0.05 and to show the co-efficients of the terms to get a since of direction.

We also quantiled the data, so we can see the difference in interaction per bracket