línea horizontal

Homework 2

A pricing model for diamond stones

**Statistical Data Analysis - 8th of December, 2019**

**Group 5**

Antoñanzas Martínez, Guillermo

Pueblas Núñez, Rodrigo

Martínez Capella, Ignacio

Burrell, David

[**Brief data description, preparation and manipulation**](#_14mpx6a8znb7)1

[**Research questions, plots and findings**](#_fmq612mgkwh3)1

[Is there a relationship between the type, genre and score?](#_x213mh3ytomu) 1

[How do the score sources differ for genre and MPAA rating?](#_aozh8cd48jsg) 2

[Genre](#_o2ge01nut8oo) 2

[MPAA rating](#_kdgqyf39wnk6) 2

[Are Oscar nominated films preferred? Which award improves the score?](#_3oj9jrwzknaw) 2

[What are the trends over the years?](#_net6iqjmtvi6) 3

[Which is the preferred month for releases? How does this change over the years?](#_8lyprlhshv4t) 3

[Do older films tend to have a higher score/number of IMDB votes?](#_wbpzuf98mki6) 4

[Do scores of movies directed by Oscar winning directors have any trend?](#_ogcf1vyov4n) 4

[**Data analysis plan**](#_n92o6gjis4o3)4

[Question 1: Is there a relationship between the type/genre and score?](#_qojjpsbnt9jm) 4

[Question 2: How do the score sources differ for genre and MPAA rating?](#_9q6qssu4xs0l) 5

[Question 3: Are oscar-awarded films more liked?](#_pgn7o2juctu0) 5

[Question 4: What are the trends over the years?](#_1lalpfyh2rg) 5

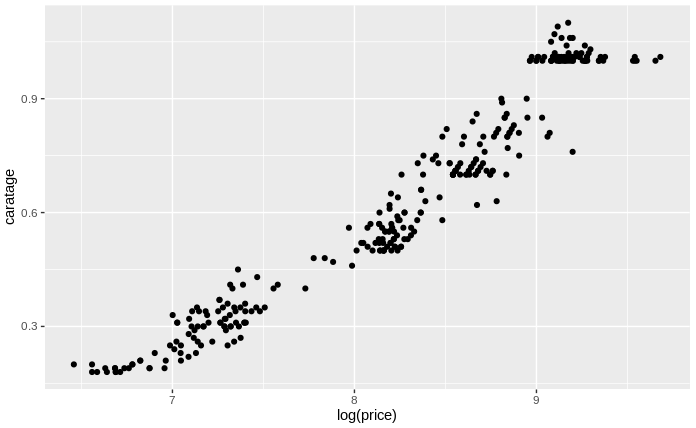
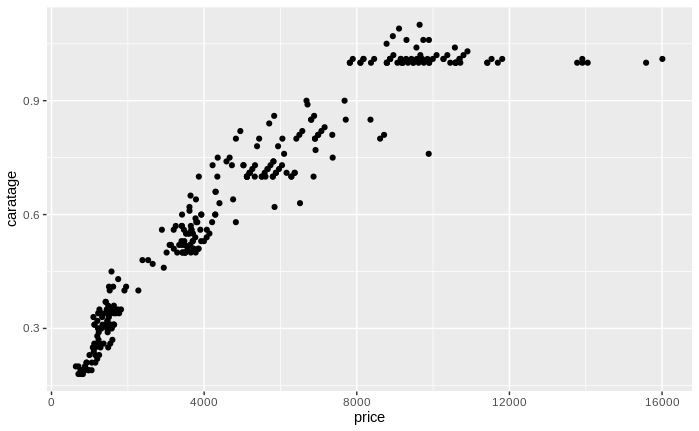
[Question 5: Do scores for movies directed by Oscar winning directors have any trend?](#_sdijy34a5mwv) 5

# Brief data description, preparation and manipulation

The data provided consists on a dataset of diamond pricing containing 308 observations listed in an advertisement, with information related to caratage, purity, clarity, institution certificate and price. According to the instructions given, we first named all the variables. Later, we ordered the purity and clarity categorical variables according to the given criterion (D, E, F, G, H, I for purity, and VS2, VS1, VVS2, VVS1, IF for clarity). No further preparation, manipulation or cleaning was needed before addressing the questions.

# Research questions, plots and findings

## Plot price vs caratage and log(price) vs caratage. Decide on which response variable is better to use.



At a quick glance we can see that the logarithmic transformation improves the normality of the data. We prefer the second plot, which follows a more linear pattern.

## Find a suitable way to include, besides caratage, the other categorical information available: clarity, color and certificate. Use the worst level of each categorical variable as the reference category and HRD for certification institution. Comment on the model fitted and perform a basic analysis of the residuals.

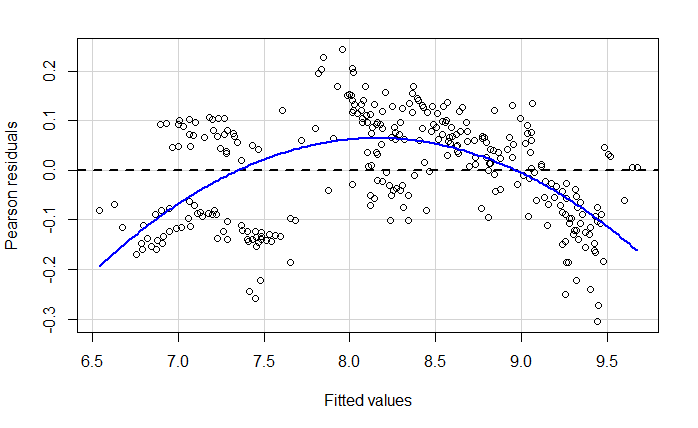
Our first approach, model 1.1, was to try to fit all the explanatory variables in a linear model with the sum of all of them: . With and , we see that we already get a good estimate on out response variable. Anyways, as we are not using the logarithmic expression yet, we get a residual standard error of 710.4.

Only changing that on our model 1.2: , we have made improvements over the first one. We achieve and , with a residual standard error of 0.1382.

We will now change, for model 1.3, the reference value on the categorical variables with the worst of each variable, as the question say. We chose as our reference value the categories I for purity, VS2 for clarity and HRD for certificate institution. However, all the results remain the same.

How could we further improve our model? For our model 1.4, we will try to add all the interactions among the variables, to see the impact of each of them in our results: . This improves even more our results, but we may be overfitting the data. This is, we are making our model too specific for our small dataset, and thus unable to better generalize to unseen data. Because of this, we will only consider the interactions with most effect, or higher p-value: caratage against certificate institution. For our final model 1.5, we then have: . With this model, we achieve and , with a residual standard error of 0.1113. All the components of our model have a p-value lower than .

To check whether the model is following the LINE assumptions for the residuals or not, we will make some tests. First of all, we are going to plot them against the fitted values:



By looking at this plot we think these residuals are not following two LINE properties: Linearity, as there is a trend along the Y axis with an inverted U shape, Equal variances, as the residuals in the left of the plot are more spread than the ones in the right. Apart from that, after performing several tests (Jarque Bera for normality, Durbin Watson for independence and Breusch Pagan test for heteroskedasticity), we see that all of them reject their respective null hypothesis, so in addition to the former analysis we can say that our model is built over failed assumptions for linear regression.

## Try two different remedial actions:

### Create a new categorical variable to segregate the stones according to caratage: let’s say less than 0.5 carats (small), less than 1 carat (medium) and 1 carat and over (large). Make small as the reference category. Add this new variable to the existing model as well as an interaction term between this new variable and caratage.

For our new model 1.5.1, we will add an explanatory variable caratage\_category according to this criterion: .

* + 1. Is this regression model satisfactory? Are the standard assumptions of linear regression validated? Are the numerical estimates sensible?
    2. Interpret the interaction parameter med\*carat. What can we infer on the incremental pricing of caratage in the 3 clusters?
    3. Which is more highly valued: colour or clarity?
    4. All other things being equal, what is the average price difference between a grade D diamond and another one graded I? And E?
    5. All other things being equal, are there price differences amongst the stones appraised by the GIA, IGI and HRD?
  1. Include the square of carat as a new explanatory variable. It avoids the subjectivity of clusters definition.

1. Which of the two remedial actions do you prefer and why? Think of terms of interpretability and validity of the assumptions.