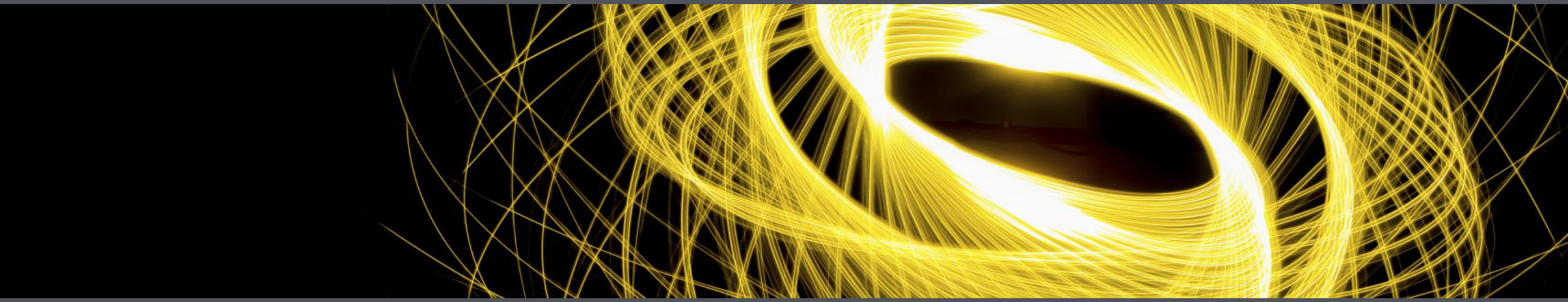
Department of Computer Science

University of Reading

CSMAD21 – Applied Data Science with Python



Exploratory Data Analysis (EDA)

University of Reading

Lecture Objectives

- Acknowledge the concept of EDA, its characteristics and importance.
- Identify different types of variables (Quantitative and Qualitative).
- Distinguish among nominal qualitative, ordinal qualitative.
- Differentiate between discrete and continuous variables.
- Acknowledge the different types of graphs and best practices.

Outline

- What is EDA and why is important?
- Characteristics of the EDA process
- Variable Types
  - Types of Qualitative Data
  - Types of Quantitative Data
- Basic EDA
- Variables and Visualisations
  - Univariate
  - Multivariate
- Summary
- Q&A

What is EDA and why is important?

- Exploratory data analysis (EDA) is used by data scientists to analyse and investigate data sets and summarize their main characteristics, often employing data visualization methods. Provides a provides a better understanding of data set variables and the relationships between them.
- Originally developed by American mathematician John Tukey in the 1970s, EDA techniques continue to be a widely used method in the data discovery process today. (IBM Cloud Education, 2020)

https://i1.wp.com/analyticsarora.com/wp-content/uploads/2021/05/Exploratory-Data-Analysis-in-Python.png?resize=800%2C600&ssl=1

What is EDA and why is important?

- The main purpose of EDA is to help look at data before making any assumptions. It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, find interesting relations among the variables.

- EDA is not a formal process with a strict set of rules. It is a state of mind. During the initial phases of EDA you should feel free to investigate every idea that occurs to you. (Hadley Wickham, 2017)

- “There are no routine statistical questions, only questionable statistical routines.” — Sir David Cox

University of Reading

Characteristics of the EDA process

It is fundamentally a creative process. And like most creative processes, the key to asking *quality* questions is to generate a large *quantity* of questions.

It is difficult to ask revealing questions at the start of your analysis because you do not know what insights are contained in your dataset.

On the other hand, each new question that you ask will expose you to a new aspect of your data and increase your chance of making a discovery.

Variable Types

- **Qualitative Variable:** A qualitative variable describes qualities or characteristics.
  - E.g. Country of origin, gender, name, or hair colour.

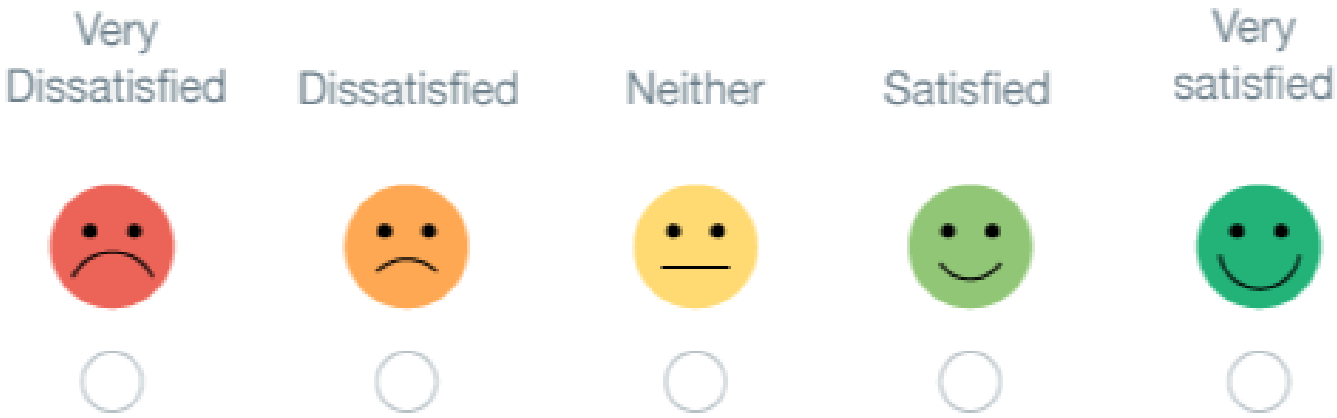
- **Quantitative Variable:** A quantitative variable addresses measurable characteristics.
  - E.g. Height, weight, or temperature.

Types of Qualitative Data

- Qualitative variables can be further classified into two types: **nominal** and **ordinal**.
  - **Nominal qualitative variables** are categories that cannot be ranked. For example, let's consider a few types of fruit: bananas, grapes, apricots, and apples. These are nominal variables because there is no implied ranked order among them. A banana, for instance, is not ranked more highly than an apricot.
    - One way to remember the definition of a nominal variable is: Nominal = Named.

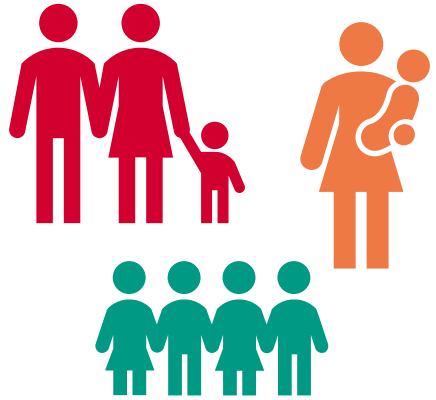
Types of Qualitative Data

- **Ordinal qualitative variables** can be ranked. They are qualitative because they are not numerically measurable, but there is a logical rank-order among them. For example, think of surveys you may have taken. Examples of ordinal qualitative values on surveys are:
  - Never, Sometimes, Mostly, Always; Extremely dissatisfied, Dissatisfied, Neither satisfied nor dissatisfied, Satisfied, Extremely satisfied
  - One way to remember the definition of an ordinal variable is: Ordinal = Ordered.



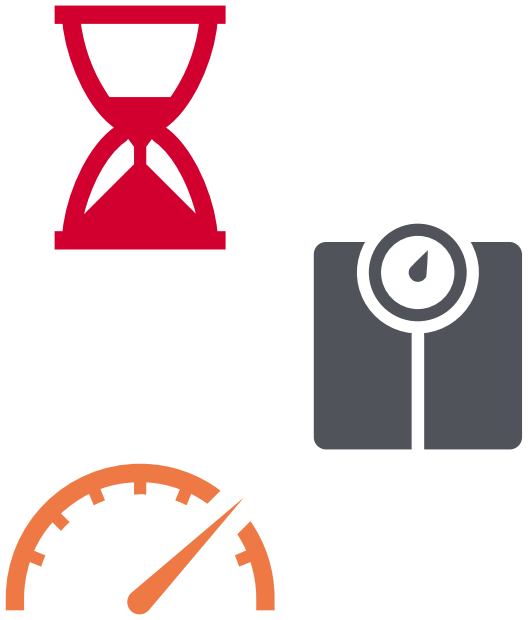
Types of Quantitative Data

- Quantitative variables can be further classified into two types: **discrete** and **continuous**.
  - **Discrete variables** are individually separate and distinct. Simply stated, **if you can count it individually, it is a discrete variable**. For example, you can count the number of children in a household individually. A household can have 0 children, 3 children, 6 children, and so on, but it can not have 3.45 children.
  - Other examples: Total number of socks in a drawer, total number of toes on all the feet of all the people in your city is even a discrete variable(It would take a long time to individually count all those toes, but it's still possible to do so).



Types of Quantitative Data

- Quantitative variables can be further classified into two types: **discrete** and **continuous**.
  - **Continuous** means forming an unbroken whole, without interruption. These are variables that cannot be counted in a finite amount of time because there is an infinite number of values between any two values. For example, if you want to measure time, every unit of time can be broken into even smaller units: The response time to a stimulus could be expressed as 1.64 seconds, or it could be further broken down and expressed as 1.642378765 seconds, and so on, infinitely.
  - Other examples of continuous values include temperature, distance, and mass.



Variables and Visualisations

- Quantitative and qualitative variables have different uses in visualizations:
  - **Quantitative** variables are the data **elements you can calculate**. They can also be aggregated (sum and average are two examples of aggregation).
  - **Qualitative** variables set the level of detail in the visualization. **They can be used to categorise**, segment, and reveal the details in your data.

Types of Data Example

	carat	cut	color	clarity	depth	table	price	x	y	z
0	0.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
1	0.21	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31
2	0.23	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31
3	0.29	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63
4	0.31	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75

Basic EDA

- Some of the basic steps in EDA are:
  - Head of the dataset
  - Shape of the dataset
  - Info of the dataset (data dictionary)
  - Summary of the dataset

```
diamonds.shape
```

```
(53940, 10)
```

	carat	cut	color	clarity	depth	table	price	x	y	z
0	0.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
1	0.21	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31
2	0.23	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31
3	0.29	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63
4	0.31	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75

Content

price price in US dollars (\$326--\$18,823)

carat weight of the diamond (0.2--5.01)

cut quality of the cut (Fair, Good, Very Good, Premium, Ideal)

color diamond colour, from J (worst) to D (best)

clarity a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))

x length in mm (0--10.74)

y width in mm (0--58.9)

z depth in mm (0--31.8)

depth total depth percentage = z / mean(x, y) = 2 * z / (x + y) (43--79)

table width of top of diamond relative to widest point (43--95)

https://www.kaggle.com/shivam2503/diamonds

```
diamonds.info()
```

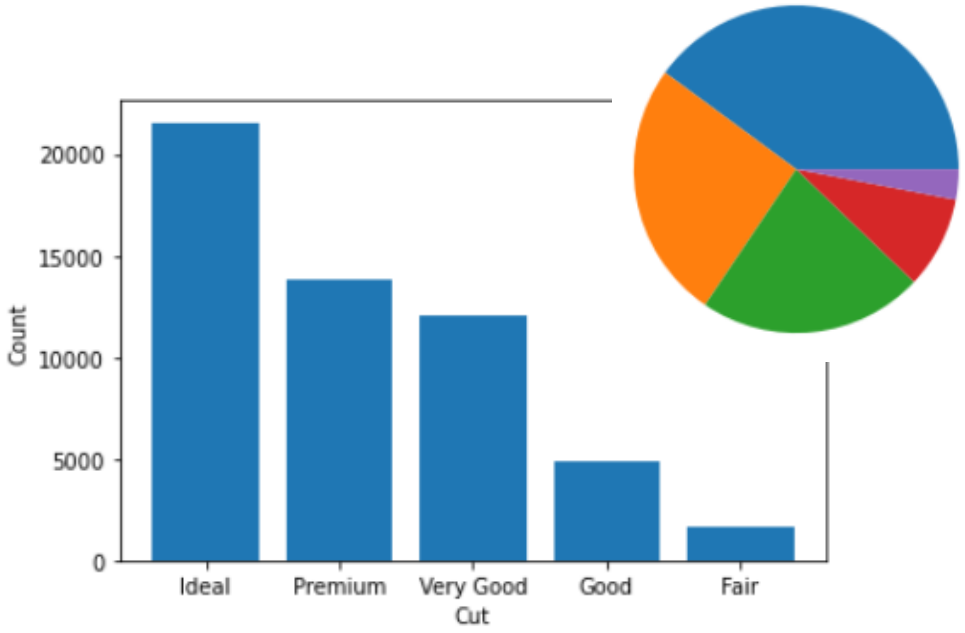
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 53940 entries, 0 to 53939
Data columns (total 10 columns):
 #   Column   Non-Null Count  Dtype  
---  ------   --------------  -----  
 0   carat    53940 non-null  float64
 1   cut      53940 non-null  object 
 2   color    53940 non-null  object 
 3   clarity  53940 non-null  object 
 4   depth    53940 non-null  float64
 5   table    53940 non-null  float64
 6   price    53940 non-null  int64  
 7   x        53940 non-null  float64
 8   y        53940 non-null  float64
 9   z        53940 non-null  float64
dtypes: float64(6), int64(1), object(3)
memory usage: 4.1+ MB
```

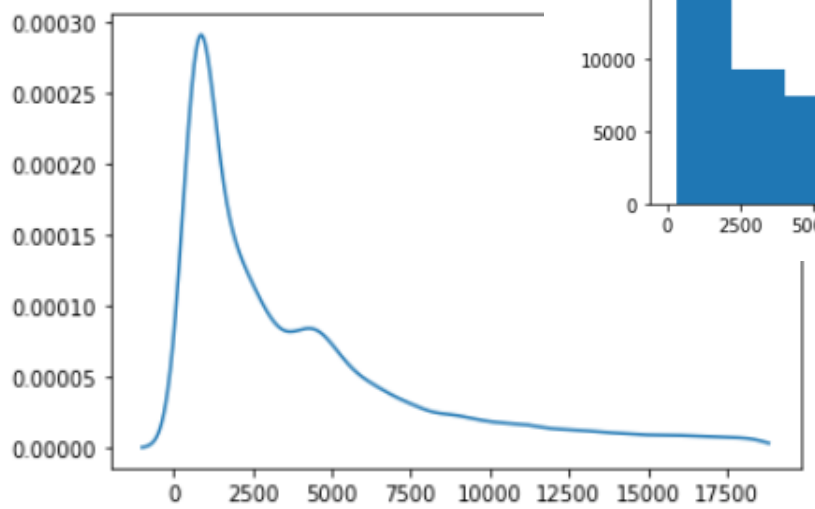
```
diamonds.describe()
```

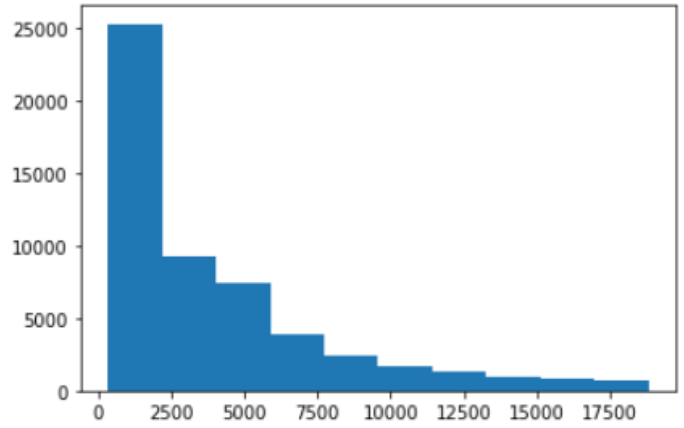
	carat	depth	table	price	x	y	z
count	53940.000000	53940.000000	53940.000000	53940.000000	53940.000000	53940.000000	53940.000000
mean	0.797940	61.749405	57.457184	3932.799722	5.731157	5.734526	3.538734
std	0.474011	1.432621	2.234491	3989.439738	1.121761	1.142135	0.705699
min	0.200000	43.000000	43.000000	326.000000	0.000000	0.000000	0.000000
25%	0.400000	61.000000	56.000000	950.000000	4.710000	4.720000	2.910000
50%	0.700000	61.800000	57.000000	2401.000000	5.700000	5.710000	3.530000
75%	1.040000	62.500000	59.000000	5324.250000	6.540000	6.540000	4.040000
max	5.010000	79.000000	95.000000	18823.000000	10.740000	58.900000	31.800000

•Variables and Visualisations – Univariate

- Common types of univariate graphics include:
  - Quantitative data: Density plots, histograms and box plots.
  - Qualitative data: Bar plot, pie chart, and variations of these graphs.

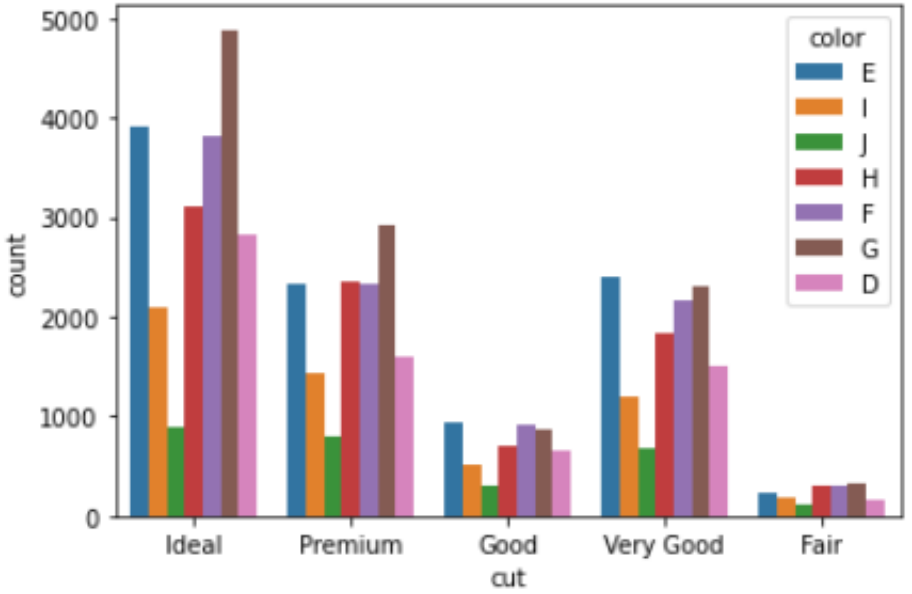






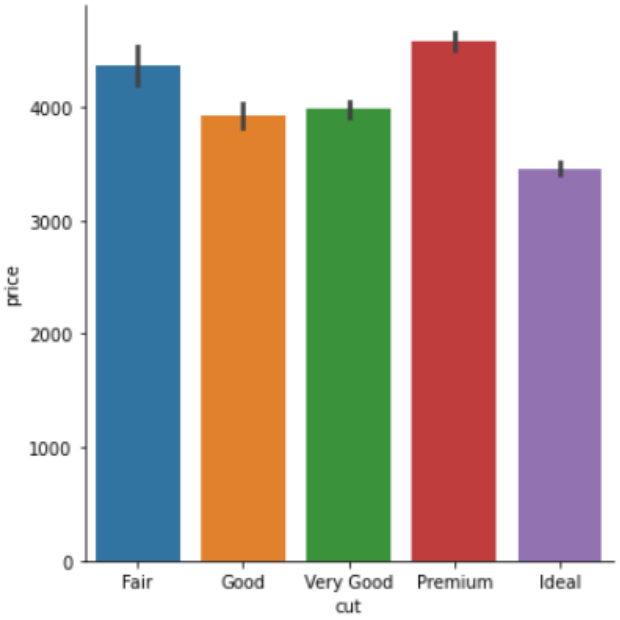
•Variables and Visualisations - Multivariate

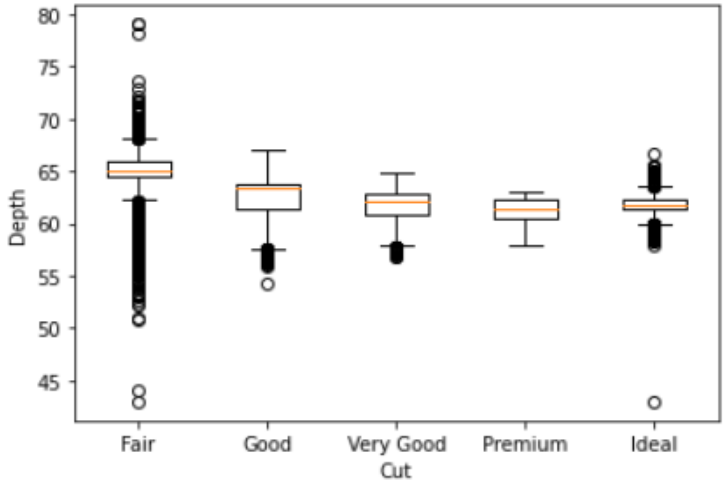
- Qualitative vs Qualitative
  - Bar chart
  - Grouped bar chart



•Variables and Visualisations - Multivariate

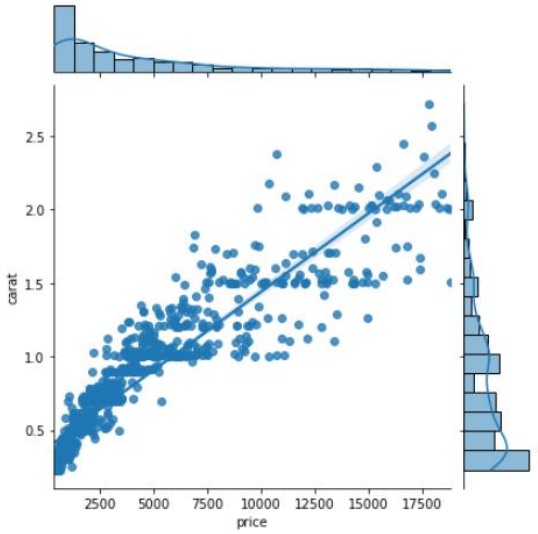
- Qualitative vs. Quantitative
  - Bar chart
  - Violin plot
  - Categorical box plot
  - Swarm plot

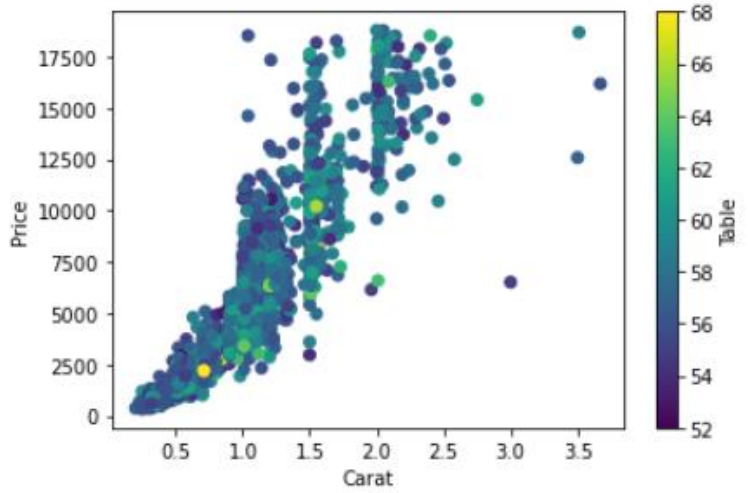


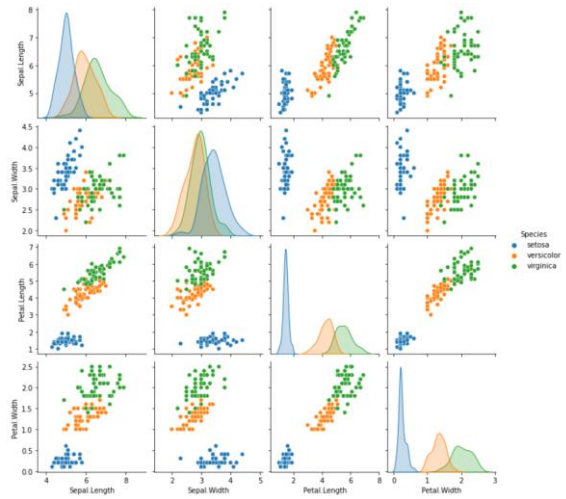


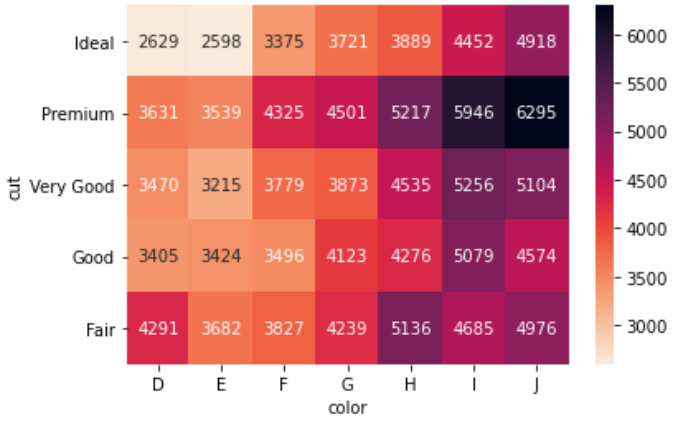
•Variables and Visualisations - Multivariate

- Quantitative vs. Quantitative
  - Scatterplot
  - Line plot
  - Heatmap for correlation
  - Joint plot









University of Reading

Summary

- Th objective of EDA is to discover the data in general and the possible relationships in it. It is a creative process where different hypothesis and ideas can be explored.
- Qualitative data can be classified in **nominal** and **ordinal** types. Nominal is data that can be defined as categories and ordinal can be ranked logically. This kind of variable can be use to categorise the data.
- Quantitative data can be classified in **discrete** and **continuous** types. Discrete are data point that can be count individually, and continuous is data that form an unbroken whole, without interruption. This can of data allows calculations and aggregations.
- The aggregation functions are sum, average, median, minimum, maximum and count. Granularity allows to see the data in more detail.
- You need to consider the data type in order to create the graph that fits better and that explains better the data.

Questions

- Let me know during the practical (face to face or online).
- Send me an email to: **m.sanchezrazo@reading.ac.uk**
- **Book a meeting with me in the following link:**

https://outlook.office365.com/owa/calendar/MiguelSanchezAppointments@livereadingac.onmicrosoft.com/bookings/

References

- Data Literacy for All (2020). Available at: https://elearning-samples.tableau.com/page/data-literacy
- Chapter 1, Hands-On Exploratory Data Analysis with Python, Suresh Kumar (2020).
- Exploratory Data Analysis(2020), IBM Cloud Education, Available at: https://www.ibm.com/cloud/learn/exploratory-data-analysis