EBU5601
Data Design

Introduction to Data Analysis

Dr Chao Shu, Dr Xiaolan Liu
School of Electronic Engineering and Computer Science
Queen Mary University of London
Sep. 2024

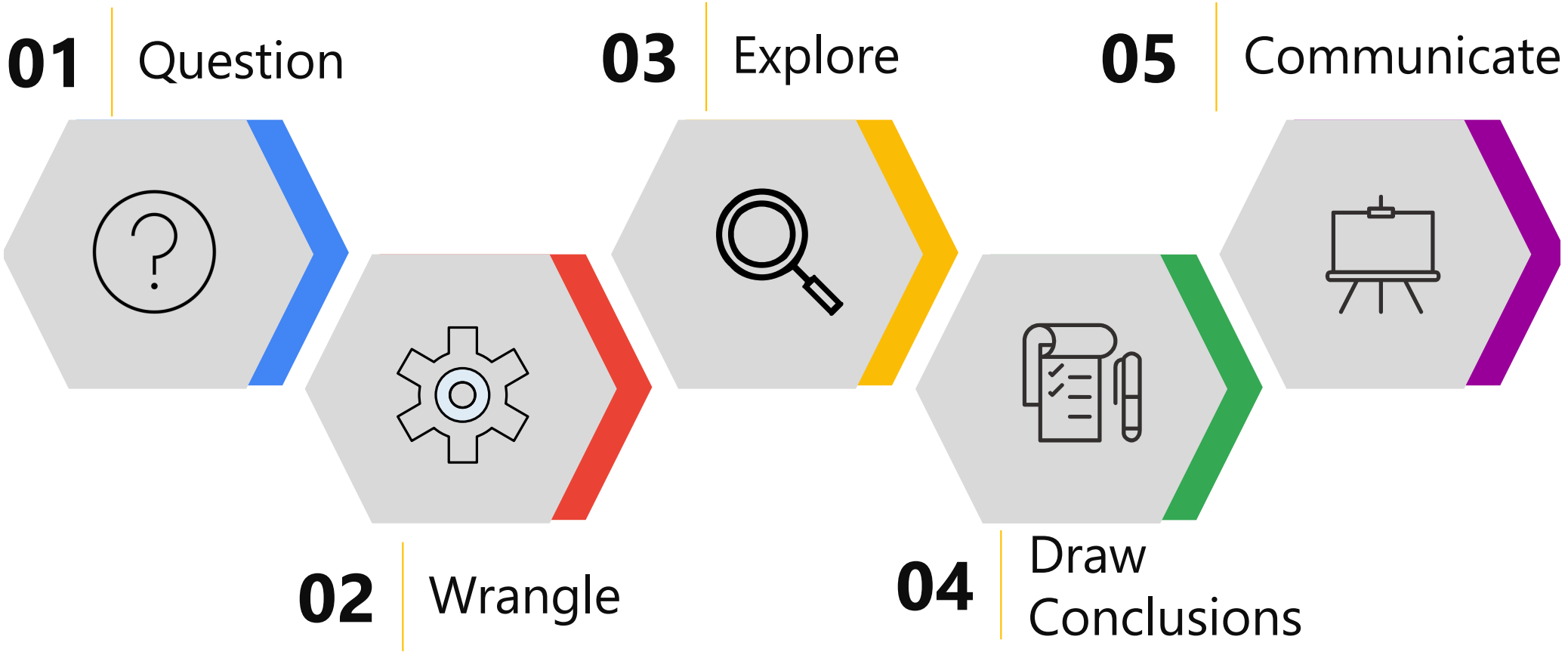
Learning Outcomes

- The main outcomes are:
  - **[LO1.1]** Understand the main steps of the data analysis process
  - **[LO1.2]** Understand the format of the comma separated values (csv) file and be able to read csv files using pandas
  - **[LO1.3]** Be familiar with the basic techniques for assessing and cleaning data using pandas

DATA ANALYSIS OVERVIEW

Data Analysis Process Overview

- The data analysis process can be organised into five steps:



Data Analysis Process Overview

- **Step 1: Ask Questions**
  - Ask questions about given data.
  - Ask questions first and gather data based on that late
- **Step 2: Wrangle Data**
  - **Gather** the data you need to answer your questions.
    - Downloading files that are readily available
    - Getting data from an API or web scraping
    - Pulling data from existing databases
    - Combine data from multiple different formats.
  - **Assess** the data to identify any problems in the data's quality or structure.
  - **Clean** the data by modifying, replacing, or removing data to ensure that your dataset is of the highest quality and as well-structured as possible.

Data Analysis Process Overview

- Comma Separated Values (csv) File
  - A text file with a tabular structure that holds only raw data. It is easy to process manually using code such as python.

datetime	season	holiday	workingday	weather	temp	atemp	humidity	windspeed	casual	registered	count
01/01/2011 00:00	1	0	0	1	9.84	14.395	81	0	3	13	16
01/01/2011 01:00	1	0	0	1	9.02	13.635	80	0	8	32	40
01/01/2011 02:00	1	0	0	1	9.02	13.635	80	0	5	27	32
01/01/2011 03:00	1	0	0	1	9.84	14.395	75	0	3	10	13
01/01/2011 04:00	1	0	0	1	9.84	14.395	75	0	0	1	1
01/01/2011 05:00	1	0	0	2	9.84	12.88	75	6.0032	0	1	1
01/01/2011 06:00	1	0	0	1	9.02	13.635	80	0	2	0	2
01/01/2011 07:00	1	0	0	1	8.2	12.88	86	0	1	2	3
01/01/2011 08:00	1	0	0	1	9.84	14.395	75	0	1	7	8
01/01/2011 09:00	1	0	0	1	13.12	17.425	76	0	8	6	14

Header

```
1  datetime,season,holiday,workingday,weather,temp,atemp,humidity,windspeed,casual,registered,count
2  01/01/2011 00:00,1,0,0,1,9.84,14.395,81,0,3,13,16
3  01/01/2011 01:00,1,0,0,1,9.02,13.635,80,0,8,32,40
4  01/01/2011 02:00,1,0,0,1,9.02,13.635,80,0,5,27,32
5  01/01/2011 03:00,1,0,0,1,9.84,14.395,75,0,3,10,13
6  01/01/2011 04:00,1,0,0,1,9.84,14.395,75,0,0,1,1
7  01/01/2011 05:00,1,0,0,2,9.84,12.88,75,6.0032,0,1,1
8  01/01/2011 06:00,1,0,0,1,9.02,13.635,80,0,2,0,2
9  01/01/2011 07:00,1,0,0,1,8.2,12.88,86,0,1,2,3
10 01/01/2011 08:00,1,0,0,1,9.84,14.395,75,0,1,7,8
11 01/01/2011 09:00,1,0,0,1,13.12,17.425,76,0,8,6,14
```

Data Analysis Process Overview

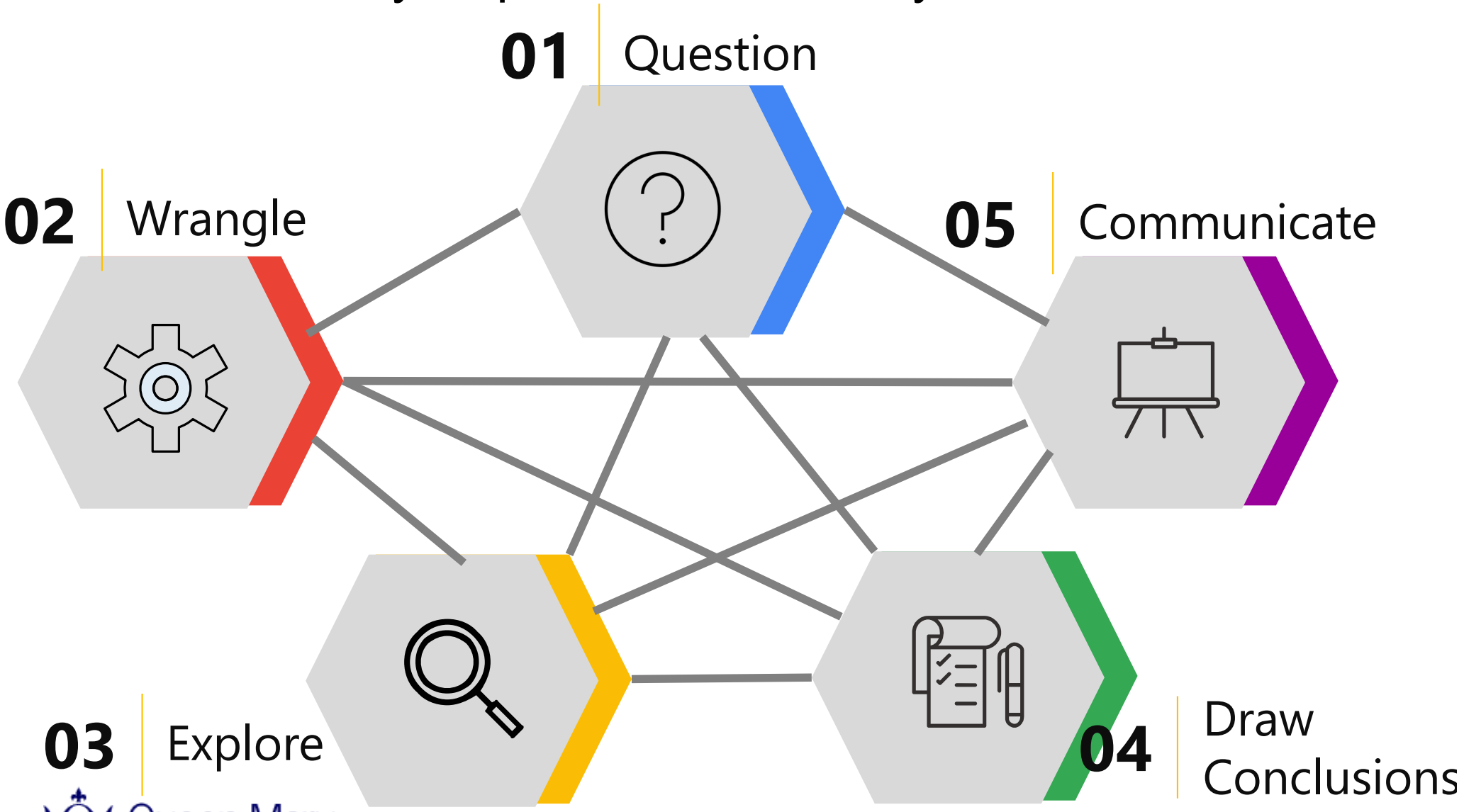
- **Step 3: Perform EDA (Exploratory Data Analysis)**
  - Explore and then augment the data to maximize the potential of your analyses, visualizations, and models.
  - Exploring involves finding patterns in your data, visualizing relationships in your data, and building intuition about what you're working with.
  - After exploring, you can do things like remove outliers and create better features from your data, also known as feature engineering.
- **Step 4: Draw conclusions (or make predictions)**
  - This step is typically approached with descriptive/inferential statistics or machine learning.
  - In this module we mainly focus on drawing conclusions with descriptive/inferential statistics.

Data Analysis Process Overview

- **Step 5: Communicate Data Findings**
  - Justify and convey meaning in the insights you've found.
  - If your end goal is to build a system, you usually need to share what you've built, explain how you reached design decisions, and report how well it performs.
  - There are many ways to communicate your results: reports, slide decks, blog posts, emails, presentations, or even conversations. Data visualization will always be very valuable.

Data Analysis Process Overview

- The data analysis process isn't always linear.



DATA ANALYSIS CASE STUDY

Data Analysis Process Demo

Demo: Bike Sharing Demand

- We'll use data from Kaggle's Bike Sharing Demand competition. (*The original dataset has been slightly modified for the purpose of this demonstration*).
- In this dataset, you are given hourly rental data spanning two years from the Capital Bikeshare program in Washington, D.C.
- In the original Kaggle competition, the participants were asked to use historical data to forecast bike rental demand.

Data Analysis Process Demo

Demo: Bike Sharing Demand

- There are 12 features in the dataset. Each feature is a column in the dataset.
- To get a decent prediction result, we must firstly get familiar with the data and conduct necessary pre-preprocess or feature engineering before we feed the data to the machine learning models.

	datetime	season	holiday	workingday	weather	temp	atemp	humidity	windspeed	casual	registered	count
0	01/01/2011 00:00	1	0	0	1	9.84	14.395	81	0.0000	3	13	16
1	01/01/2011 01:00	1	0	0	1	9.02	13.635	80	0.0000	8	32	40
2	01/01/2011 02:00	1	0	0	1	9.02	13.635	80	0.0000	5	27	32
3	01/01/2011 03:00	1	0	0	1	9.84	14.395	75	0.0000	3	10	13
4	01/01/2011 04:00	1	0	0	1	9.84	14.395	75	0.0000	0	1	1
...	...	...	...	...	...	...	...	...	...	...	...	...
10881	19/12/2012 19:00	4	0	1	1	15.58	19.695	50	26.0027	7	329	336
10882	19/12/2012 20:00	4	0	1	1	14.76	17.425	57	15.0013	10	231	241
10883	19/12/2012 21:00	4	0	1	1	13.94	15.910	61	15.0013	4	164	168
10884	19/12/2012 22:00	4	0	1	1	13.94	17.425	61	6.0032	12	117	129
10885	19/12/2012 23:00	4	0	1	1	13.12	16.665	66	8.9981	4	84	88

10886 rows × 12 columns

Data Analysis Process Demo

Demo: Bike Sharing Demand

- **Step 1: Ask Questions**

Q1: Given the above data on features that potentially influence the number of bikes rented each hour, what questions would be relevant to ask?

Go to
www.menti.com
Enter the code
5806 9241

Or use QR code

	datetime	season	holiday	workingday	weather	temp	atemp	humidity	windspeed	casual	registered	count
0	01/01/2011 00:00	1	0	0	1	9.84	14.395	81	0.0000	3	13	16
1	01/01/2011 01:00	1	0	0	1	9.02	13.635	80	0.0000	8	32	40
2	01/01/2011 02:00	1	0	0	1	9.02	13.635	80	0.0000	5	27	32
3	01/01/2011 03:00	1	0	0	1	9.84	14.395	75	0.0000	3	10	13
4	01/01/2011 04:00	1	0	0	1	9.84	14.395	75	0.0000	0	1	1
...	...	...	...	...	...	...	...	...	...	...	...	...
10881	19/12/2012 19:00	4	0	1	1	15.58	19.695	50	26.0027	7	329	336
10882	19/12/2012 20:00	4	0	1	1	14.76	17.425	57	15.0013	10	231	241
10883	19/12/2012 21:00	4	0	1	1	13.94	15.910	61	15.0013	4	164	168
10884	19/12/2012 22:00	4	0	1	1	13.94	17.425	61	6.0032	12	117	129
10885	19/12/2012 23:00	4	0	1	1	13.12	16.665	66	8.9981	4	84	88

10886 rows × 12 columns

Data Analysis Process Demo

Demo: Bike Sharing Demand

- **Step 2: Data Wrangling**

Q2: What potential problems do you see with this dataset that would need to be fixed before continuing with analysis?

Go to
www.menti.com
Enter the code
5806 9241

Or use QR code

```
Data columns (total 12 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   datetime    10886 non-null  object
 1   season      10886 non-null  int64
 2   holiday     10886 non-null  int64
 3   workingday  10886 non-null  int64
 4   weather     10886 non-null  int64
 5   temp        10848 non-null  float64
 6   atemp       10848 non-null  float64
 7   humidity    10886 non-null  int64
 8   windspeed   10886 non-null  float64
 9   casual      10886 non-null  int64
 10  registered  10886 non-null  int64
 11  count       10886 non-null  int64
```

	season	holiday	workingday	weather	temp	atemp	humidity	windspeed	casual	register
count	10886.000000	10886.000000	10886.000000	10886.000000	10848.000000	10848.000000	10886.000000	10886.000000	10886.000000	10886.0000
mean	2.506614	0.028569	0.680875	1.418427	20.244801	23.674888	61.886460	12.799395	36.021955	155.5521
std	1.116174	0.166599	0.466159	0.633839	7.991858	8.891435	19.245033	8.164537	49.960477	151.0390
min	1.000000	0.000000	0.000000	1.000000	0.820000	0.760000	0.000000	0.000000	0.000000	0.0000
25%	2.000000	0.000000	0.000000	1.000000	13.940000	16.665000	47.000000	7.001500	4.000000	36.0000
50%	3.000000	0.000000	1.000000	1.000000	20.500000	24.240000	62.000000	12.998000	17.000000	118.0000
75%	4.000000	0.000000	1.000000	2.000000	26.240000	31.060000	77.000000	16.997900	49.000000	222.0000
max	4.000000	1.000000	1.000000	4.000000	200.000000	300.000000	100.000000	56.996900	367.000000	886.0000

Data Analysis Process Demo

Demo: Bike Sharing Demand

- **Step 2: Data Wrangling**

Deal with those missing values and apparently wrong values.

1. Remove the records/rows with missing values if not too many records/rows contain missing values or wrong values.

2. Fill in all missing values with some values we think might be close to the actual values, e.g., the average temperature of the days/hours around the day/hour that do not have temperature readings. We can also do the same thing to the records with unrealistic temperature readings.

Data Analysis Process Demo

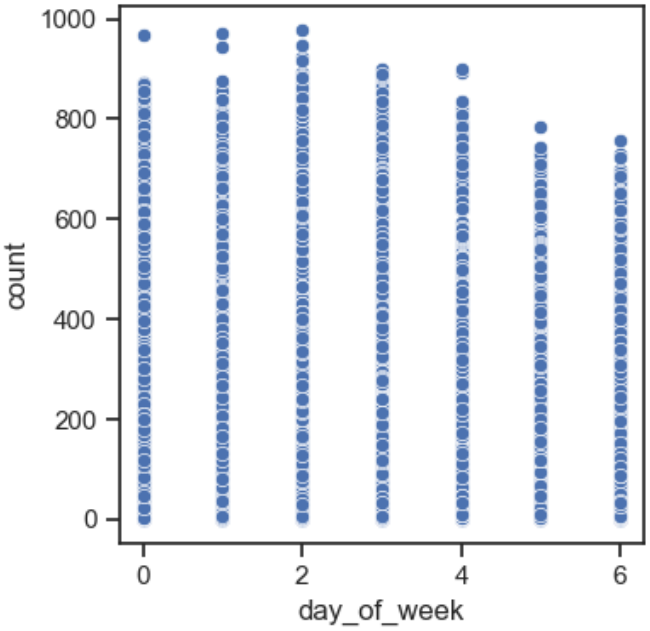
Demo: Bike Sharing Demand

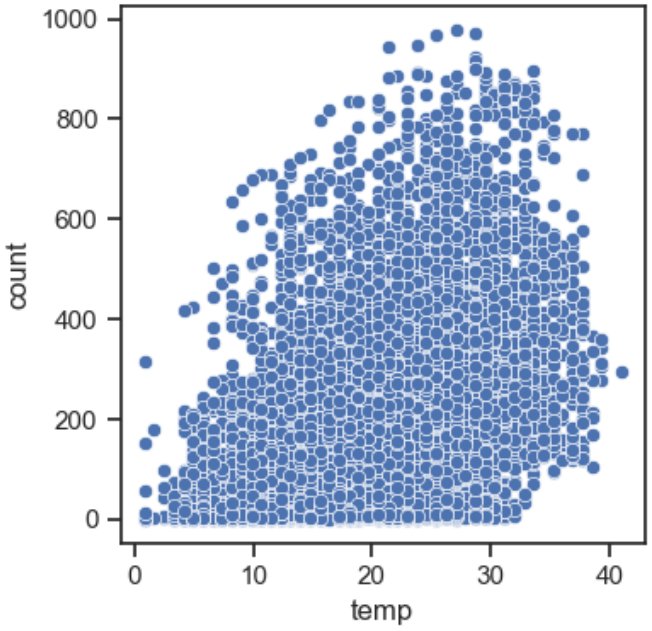
- **Step 3: Exploratory Data Analysis**

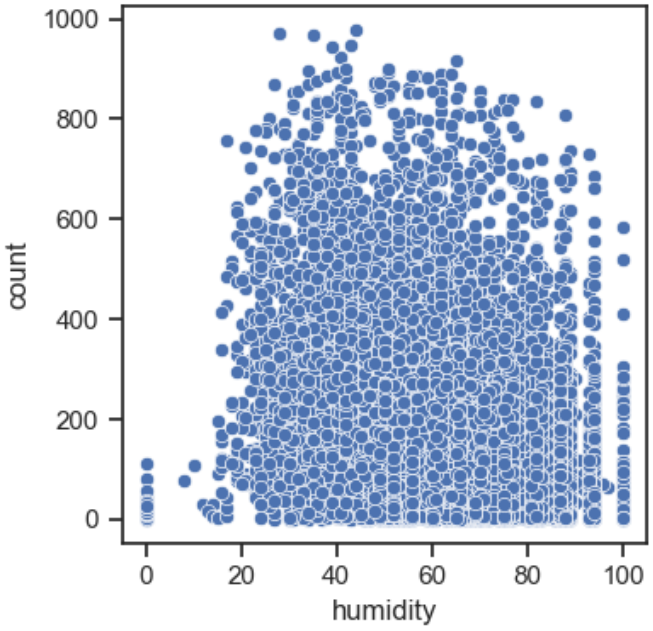
Q3: Based on these scatterplots, which of these three features seems most helpful in predicting rental count?

Go to
www.menti.com
Enter the code
5806 9241

Or use QR code





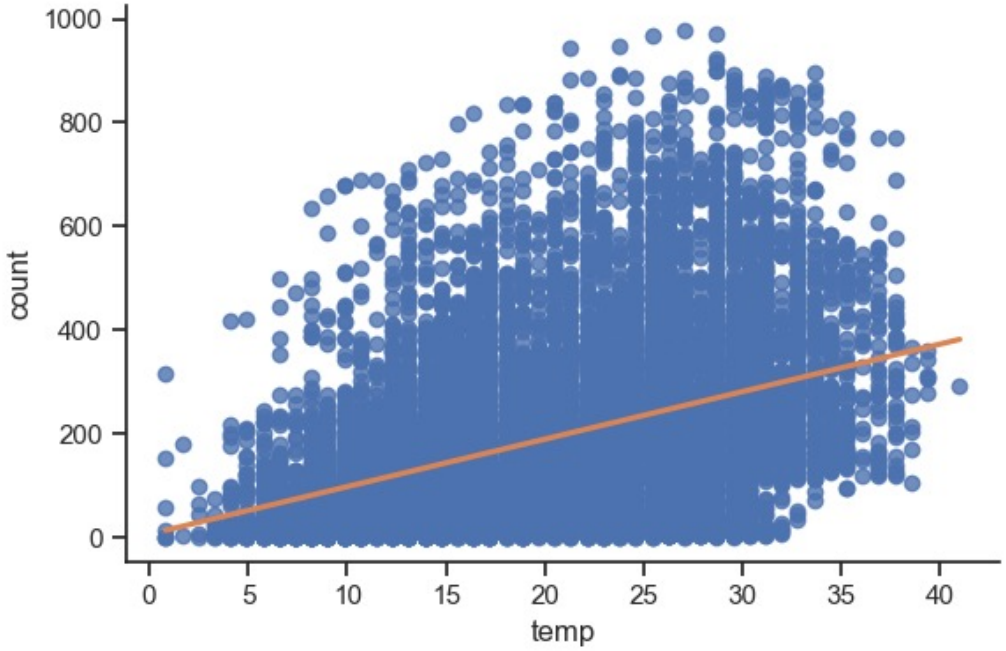


Data Analysis Process Demo

Demo: Bike Sharing Demand

- **Step 4: Draw Conclusions**

Q4: Based on this graph of regressing bike rental count on the temperature, how many additional bikes do you think would be checked out if the temperature rose from 2℃ to 30 ℃?



Go to
www.menti.com
Enter the code

5806 9241

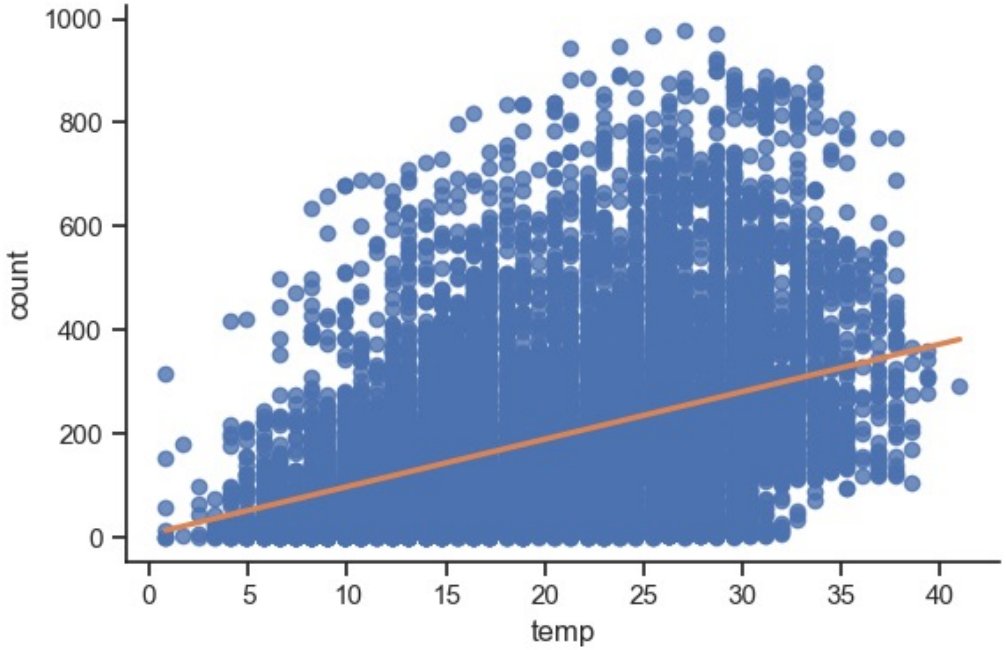
Or use QR code

Data Analysis Process Demo

Demo: Bike Sharing Demand

- **Step 5: Communicate Data Findings**

Q5: What would be valid methods of communicating your conclusions from the Bike Sharing data?



Go to
www.menti.com
Enter the code

5806 9241

Or use QR code

RECAP

Introduction to Data Analysis

Recap

- **Five Steps for Data Analysis**
  - Ask Questions
  - Wrangle Data
  - Perform Exploratory Data Analysis (EDA)
  - Draw Conclusions
  - Communicate Data Findings
- **Comma Separated Values (csv) File**
  - Format
- **Techniques**
  - Read csv file
  - Deal with missing values and anomalies

Questions

Use student forum on QM+

chao.shu@qmul.ac.uk
xiaolanliu@qmul.ac.uk