Please follow the instructions stated in this document during the Greate Research team IA period, Thanks!

Greate x HKIIT

Research theme IA Guidelines

HKIIT

Greate(HK) Limited

Table of Content

Introduction

Welcome to the Industrial Attachment (IA) program, an internship initiative initiated by HKIIT in collaboration with our esteemed IA partner, Greate (HK) Limited. This IA program focuses on providing students with valuable hands-on experience in the field of Big Data Analysis. We are proud to cooperate with Greate (HK) Limited, a renowned organization specializing in Big Data Analysis, whose expertise and guidance will be instrumental in making this IA program a success. The IA program will be conducted throughout the month of August, allowing students to fully immerse themselves in the project and accumulate a minimum of 100 working hours. Students will have the opportunity to select from a range of projects within the field of Big Data Analysis, tailoring their experience to align with their interests and learning objectives. Through this IA program, we aim to empower students, foster industry-academic collaboration, and contribute to the advancement of Big Data Analysis.

Bear in mind

Industrial Attachment (IA) Program Guideline

1. Group Formation and Project Selection:
    I. Students are required to form groups of three or four and choose one project to work on collectively.
    II. Detailed project information can be found in the attached appendix.
2. Attendance and Reporting Time:
    I. Students must return to the campus promptly at 10:00 am and mark attendance daily.
3. Working Basic Informaiton:
    I. The IA should start on **1st August 2025**, and terminate on **31st August 2025**, the **last day** depends on student working hours and performance. We would consider to early finish the IA if student have good discipline and performance during the IA.
    II. The working hours for the IA program are from 9:30 am to 5:30 pm, Monday to Friday. You have to finish your work in **Room TYC511**.
    III. Students are expected to adhere to the provided schedule in this document, ensuring punctuality for all activities.
4. Absence or Lateness Notification:
    I. If a student plans to be absent or will be late, they must inform their workplace supervisor one day in advance.
5. Final Analytical Product:
    I. The final deliverable for the IA program should be a Tableau workbook, showcasing the results of the data analysis conducted during the project.
6. Data and Information Resources:
    I. All the necessary data required for the IA program has already been posted on Moodle.
    II. Students should access the provided resources to gather the relevant data for their projects.
7. IA Workbook:
    I. All students are required to complete the IA workbook, ensuring that all necessary information is filled in accurately.
8. Workplace Supervisors:
    I. The IA program workplace supervisors are **Benson Lau** and **Blue Ho**. Students should reach out to them for guidance, support, and feedback throughout the program.

9. The following is the student list of this IA:

Name	Team	Case
NG Ping Yam	--	--
CHENG Kong Sang	1	Case 2021
CHEUNG Wang Kwong	1	Case 2021
KEUNG Pui Shing	1	Case 2021
CHENG Chun Hei	2	Case 2223
LUI Kwok Fung	2	Case 2223
NG Wang Hon	2	Case 2223
AQSA-AHMED	3	Case2324
LUI Wing Sing	3	Case2324
TSANG Man Chiu	3	Case2324
HE Xu	4	Case 2021
KUANG Chin Sing	4	Case 2021
LAM Ho Ching	4	Case 2021
GAN Zechun	5	Case 2526
ZHANG Shuang	5	Case 2526
ZHANG Weien	5	Case 2526

Please be sure to review this guideline thoroughly and reach out to your workplace supervisors or program coordinators if you have any further questions or need clarification. We wish you a successful and rewarding experience during the IA program!

Schedule

You have to follow the below schedules:

Date	To-Do list	Accumulated working hour
1st August 2025	Form Group and Project selections Project start Project Briefing Meeting Fill in the information in the IA Handbook	8
4th August 2025	Report the Project Planning Project On Progress	16
Date	To-Do list	Accumulated working hour
5th – 8th August 2024	Project On Progress Preliminary view of data Progress Report Meeting on 8th or 9th August	56
14th – 18th August 2024	Project On Progress Advanced analysis of the project (dashboarding) Progress Report Meeting on 15th or 16th August	96
21th August 2024	Final Presentation Fill in the information in the IA Handbook and submit to Moodle submission box	104

Project Descriptions

There are six projects that are across various industries and data. All Project cases have to retrieve knowledge from the data and presentation to supervisor. The Final product of each project should be a Tableau Public workbook with a ".twbx" file extension.

Project Case 2021

You are required to conduct a case study on COVID-19 in Hong Kong. you may have some questions for your starting point, the following are only some examples:

1. Any distinguishable keyword for each district? What Message bring out?
2. Overview of patient cases
3. What are the differences between two peaks of explosion?

Project Case 2122

The data is about the transaction data of British Property. All groups and individuals use the provided data and finish the following tasks: Data preparation, Mandatory task, Advanced technique task, and Finding demonstrations. All tasks and demonstrations are required to utilize the features of Tableau Public and R for data pre-processing.

To Conduct the project smoothly, you may have some questions for your starting point, and the following question is only an example:

1. Which type of property is popular in the UK?
2. How is the pandemic affecting property transactions?
3. Which kind of property will become the next trend?

Project Case 2223

With the development of information technology, people's recreation is highly reliant on social media platforms. A huge amount of user behaviour records are generated on the platforms. Therefore, understanding the platform user's behaviour can effectively promote your product to the target customers/audience.

You are required to form a group of 2~ 3 classmates or even work individually. Each group or individual will receive one set of data. All groups or individuals have to process the dataset and generate some findings. The data is about Youtube Video data from various countries. All groups and individuals use provided data.

To Conduct the project smoothly, you may have some questions for your starting point, and the following question is only an **example:**

1. Comparison between two regions of the video?
2. What is the audience interested in?
3. What is Youtube used for?
4. Is virtual Youtube popular in the western world?

Project Case 2324

The data is about LA Crime Data in the past 13 years. All groups and individuals use the provided data. To Conduct the project smoothly, you may have some questions for your starting point, and the following question is only an **example:**

1. What are the Crime Change in these 10 Years?
2. A Safety Instruction for tourists?
3. Which area is avoid?
4. What are the circumstances of crime in LA?

Project Case 2425

The data is about the transaction data of British Property. All groups and individuals use the provided data and finish the following tasks: Data preparation, Mandatory task, Advanced technique task, and Finding demonstrations. All tasks and demonstrations are required to utilise the features of Tableau Public and R for data pre-processing.

To Conduct the project smoothly, The project topic should not be limited to the following suggestions but should included in the theme of "**When Britain parallel with Pandemic**"**:**

1. Which type of property is popular in the UK during the Pandemic?
2. How is the pandemic affecting property transactions?
3. Which kind of property will become the next trend?

Project Case 2526

The data is about the Top 50 Popular Songs in 73 Countries. All tasks and demonstrations are required to utilise the features of Tableau Public and R for data pre-processing.

To Conduct the project smoothly, The project topic should not be limited to the following suggestions but should included in the theme of “Music with You”:

1. Review of song style in different country?
2. Relatship between artist and song?
3. Which kind of song will become the next trend?
4. The style of different artists.

Appendix

Project Case 2021 Full Document

You are required to form a group of 2~ 3 classmates, or even work individually. Each group will receive one set of data. All of group or individuals have to process the dataset and generate some findings. The data is about case data of COVID-19 in Hong Kong. All groups and individuals make use of the provided data and finish the following tasks: Data preparation, Mandatory task, Advanced technique task, and Finding demonstrations. All tasks and demonstrations are required to utilize the feature of Tableau desktop, or public, for data process. Extra bonus on using R on data processing.

To Conduct the project smoothly, you may have some questions for your starting point, the following are only some **examples:**

1. Any distinguishable keyword for each district? What Message brings out?
2. Overview of patient cases
3. What are the differences between two peaks of explosion?

Dataset Description

Case Dataset from Government and Open data

This is public data acquired by ESHF and CMHO Media Awareness Group. The data contain the details of each COVID-19 patient case, from January 2020 to August 2020. The dataset also included relevant articles for each case. The following is a sample of the dataset.

case_no	onset_date	confirmati	gender	age	hospital_e	status	status_en	type_en	residnecy	residency_	citizenship	citizenship	classificati	classificati	detail_en	source_url_1
1	21/1/2020	23/1/2020	M	39	Princess M	discharged	Discharge	Confirmec	non_HK	Non-HK r	Wuhan	Outside H	imported	Imported	The patien	https://www.
2	18/1/2020	23/1/2020	M	56	Princess M	discharged	Discharge	Confirmec	HK	HK reside	Ma On Sh	Sha Tin	imported	Imported	The patien	https://hd.sthe
3	20/1/2020	24/1/2020	F	62	Princess M	discharged	Discharge	Confirmec	non_HK	Non-HK r	Wuhan	Outside H	imported	Imported	The femal	https://news.r
4	23/1/2020	24/1/2020	F	62	Princess M	discharged	Discharge	Confirmec	non_HK	Non-HK r	Wuhan	Outside H	imported	Imported	The 62 ye	https://www.
5	23/1/2020	24/1/2020	M	63	Princess M	discharged	Discharge	Confirmec	non_HK	Non-HK r	Wuhan	Outside H	imported	Imported	The 63 ye	https://www.
6	21/1/2020	26/1/2020	M	47	Princess M	discharged	Discharge	Confirmec	HK	HK reside	North Poir	Eastern	imported	Imported	The patien	https://hk.app
7	21/1/2020	26/1/2020	F	68	Princess M	discharged	Discharge	Confirmec	HK	HK reside	Shenzhen	Outside H	imported	Imported	The 68 ye	https://www.
8	25/1/2020	26/1/2020	M	64	Princess M	discharged	Discharge	Confirmec	non_HK	Non-HK r	Wuhan	Outside H	imported	Imported	The 64 ye	https://www.
9	25/1/2020	29/1/2020	F	73	Princess M	discharged	Discharge	Confirmec	non_HK	Non-HK r	Wuhan	Outside H	imported	Imported	The femal	https://news.r
10	25/1/2020	29/1/2020	M	72	Princess M	discharged	Discharge	Confirmec	non_HK	Non-HK r	Wuhan	Outside H	imported	Imported	The male	https://news.r

Task

Data preparation

You will receive 4481 case for your project, you have to ETL before pour the data to Tableau:

1. Manually, or use R/Python code, Split all articles as Keyword
2. Find all the latitude and longitude data of patient reported place (citizenship_en)

Hint:

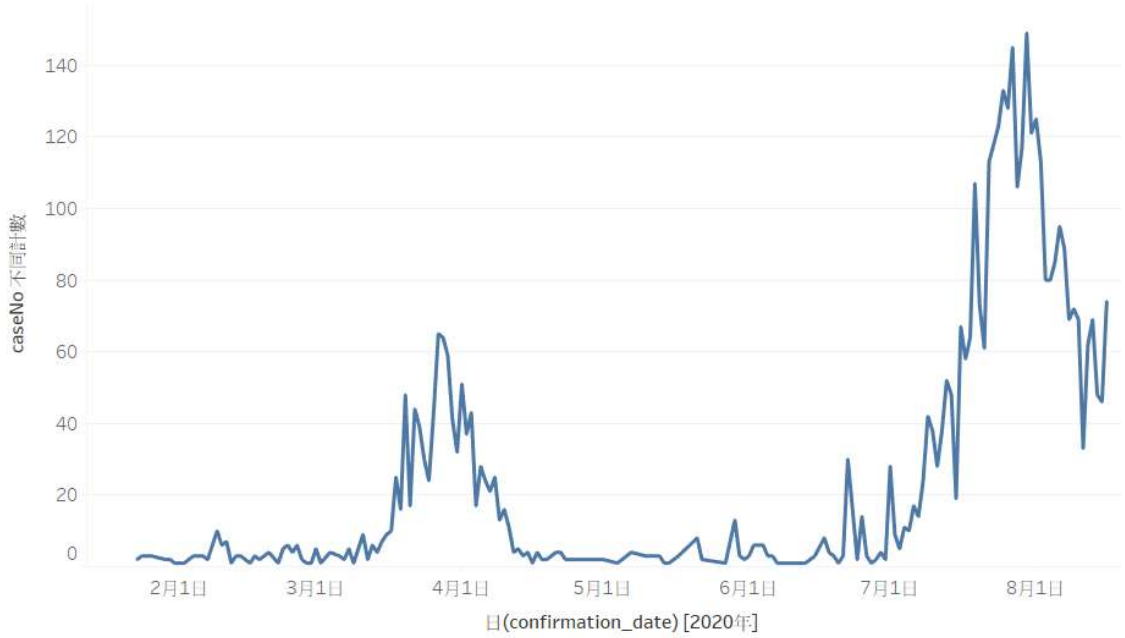
1. Filtrate all unnecessary keywords (up to you)
2. The below is just for your reference (in xlxs format)

caseNo	split	Freq	case_no	onset_date	confirmati	gender	age	hospital_e	status	status_en	type_en	residnecy	residency_	citizenship	citizenship	classificati	classificati	detail_en	source_url_1
1	21	1	1	21/1/2020	23/1/2020	M	39	Princess M	discharged	Discharge	Confirmec	non_HK	Non-HK r	Wuhan	Outside H	imported	Imported	The patien	https://www.th
1	22	1	1	21/1/2020	23/1/2020	M	39	Princess M	discharged	Discharge	Confirmec	non_HK	Non-HK r	Wuhan	Outside H	imported	Imported	The patien	https://www.th
1	a	2	1	21/1/2020	23/1/2020	M	39	Princess M	discharged	Discharge	Confirmec	non_HK	Non-HK r	Wuhan	Outside H	imported	Imported	The patien	https://www.th
1	and	4	1	21/1/2020	23/1/2020	M	39	Princess M	discharged	Discharge	Confirmec	non_HK	Non-HK r	Wuhan	Outside H	imported	Imported	The patien	https://www.th
1	at	1	1	21/1/2020	23/1/2020	M	39	Princess M	discharged	Discharge	Confirmec	non_HK	Non-HK r	Wuhan	Outside H	imported	Imported	The patien	https://www.th
1	being	1	1	21/1/2020	23/1/2020	M	39	Princess M	discharged	Discharge	Confirmec	non_HK	Non-HK r	Wuhan	Outside H	imported	Imported	The patien	https://www.th
1	blocked	1	1	21/1/2020	23/1/2020	M	39	Princess M	discharged	Discharge	Confirmec	non_HK	Non-HK r	Wuhan	Outside H	imported	Imported	The patien	https://www.th
1	detected	1	1	21/1/2020	23/1/2020	M	39	Princess M	discharged	Discharge	Confirmec	non_HK	Non-HK r	Wuhan	Outside H	imported	Imported	The patien	https://www.th
1	developed	1	1	21/1/2020	23/1/2020	M	39	Princess M	discharged	Discharge	Confirmec	non_HK	Non-HK r	Wuhan	Outside H	imported	Imported	The patien	https://www.th
1	empire	1	1	21/1/2020	23/1/2020	M	39	Princess M	discharged	Discharge	Confirmec	non_HK	Non-HK r	Wuhan	Outside H	imported	Imported	The patien	https://www.th
1	family	1	1	21/1/2020	23/1/2020	M	39	Princess M	discharged	Discharge	Confirmec	non_HK	Non-HK r	Wuhan	Outside H	imported	Imported	The patien	https://www.th
1	fever	2	1	21/1/2020	23/1/2020	M	39	Princess M	discharged	Discharge	Confirmec	non_HK	Non-HK r	Wuhan	Outside H	imported	Imported	The patien	https://www.th
1	flight	1	1	21/1/2020	23/1/2020	M	39	Princess M	discharged	Discharge	Confirmec	non_HK	Non-HK r	Wuhan	Outside H	imported	Imported	The patien	https://www.th
1	four	1	1	21/1/2020	23/1/2020	M	39	Princess M	discharged	Discharge	Confirmec	non_HK	Non-HK r	Wuhan	Outside H	imported	Imported	The patien	https://www.th
1	high	1	1	21/1/2020	23/1/2020	M	39	Princess M	discharged	Discharge	Confirmec	non_HK	Non-HK r	Wuhan	Outside H	imported	Imported	The patien	https://www.th
1	hong	1	1	21/1/2020	23/1/2020	M	39	Princess M	discharged	Discharge	Confirmec	non_HK	Non-HK r	Wuhan	Outside H	imported	Imported	The patien	https://www.th
1	hospital	1	1	21/1/2020	23/1/2020	M	39	Princess M	discharged	Discharge	Confirmec	non_HK	Non-HK r	Wuhan	Outside H	imported	Imported	The patien	https://www.th
2	travelled	1	2	18/1/2020	23/1/2020	M	56	Princess M	discharged	Discharge	Confirmec	HK	HK reside	Ma On Sh	Sha Tin	imported	Imported	The patien	ht
2	treatment	2	2	18/1/2020	23/1/2020	M	56	Princess M	discharged	Discharge	Confirmec	HK	HK reside	Ma On Sh	Sha Tin	imported	Imported	The patien	ht
2	wales	1	2	18/1/2020	23/1/2020	M	56	Princess M	discharged	Discharge	Confirmec	HK	HK reside	Ma On Sh	Sha Tin	imported	Imported	The patien	ht
2	was	1	2	18/1/2020	23/1/2020	M	56	Princess M	discharged	Discharge	Confirmec	HK	HK reside	Ma On Sh	Sha Tin	imported	Imported	The patien	ht
2	wu	1	2	18/1/2020	23/1/2020	M	56	Princess M	discharged	Discharge	Confirmec	HK	HK reside	Ma On Sh	Sha Tin	imported	Imported	The patien	ht
2	wuhan	2	2	18/1/2020	23/1/2020	M	56	Princess M	discharged	Discharge	Confirmec	HK	HK reside	Ma On Sh	Sha Tin	imported	Imported	The patien	ht
3	19	1	3	20/1/2020	24/1/2020	F	62	Princess M	discharged	Discharge	Confirmec	non_HK	Non-HK r	Wuhan	Outside H	imported	Imported	The femal	ht
3	20	1	3	20/1/2020	24/1/2020	F	62	Princess M	discharged	Discharge	Confirmec	non_HK	Non-HK r	Wuhan	Outside H	imported	Imported	The femal	ht
3	23	1	3	20/1/2020	24/1/2020	F	62	Princess M	discharged	Discharge	Confirmec	non_HK	Non-HK r	Wuhan	Outside H	imported	Imported	The femal	ht
3	a	4	3	20/1/2020	24/1/2020	F	62	Princess M	discharged	Discharge	Confirmec	non_HK	Non-HK r	Wuhan	Outside H	imported	Imported	The femal	ht
3	almost	1	3	20/1/2020	24/1/2020	F	62	Princess M	discharged	Discharge	Confirmec	non_HK	Non-HK r	Wuhan	Outside H	imported	Imported	The femal	ht
3	and	3	3	20/1/2020	24/1/2020	F	62	Princess M	discharged	Discharge	Confirmec	non_HK	Non-HK r	Wuhan	Outside H	imported	Imported	The femal	ht
3	animals	1	3	20/1/2020	24/1/2020	F	62	Princess M	discharged	Discharge	Confirmec	non_HK	Non-HK r	Wuhan	Outside H	imported	Imported	The femal	ht
3	any	1	3	20/1/2020	24/1/2020	F	62	Princess M	discharged	Discharge	Confirmec	non_HK	Non-HK r	Wuhan	Outside H	imported	Imported	The femal	ht
3	appeared	1	3	20/1/2020	24/1/2020	F	62	Princess M	discharged	Discharge	Confirmec	non_HK	Non-HK r	Wuhan	Outside H	imported	Imported	The femal	ht
3	but	1	3	20/1/2020	24/1/2020	F	62	Princess M	discharged	Discharge	Confirmec	non_HK	Non-HK r	Wuhan	Outside H	imported	Imported	The femal	ht

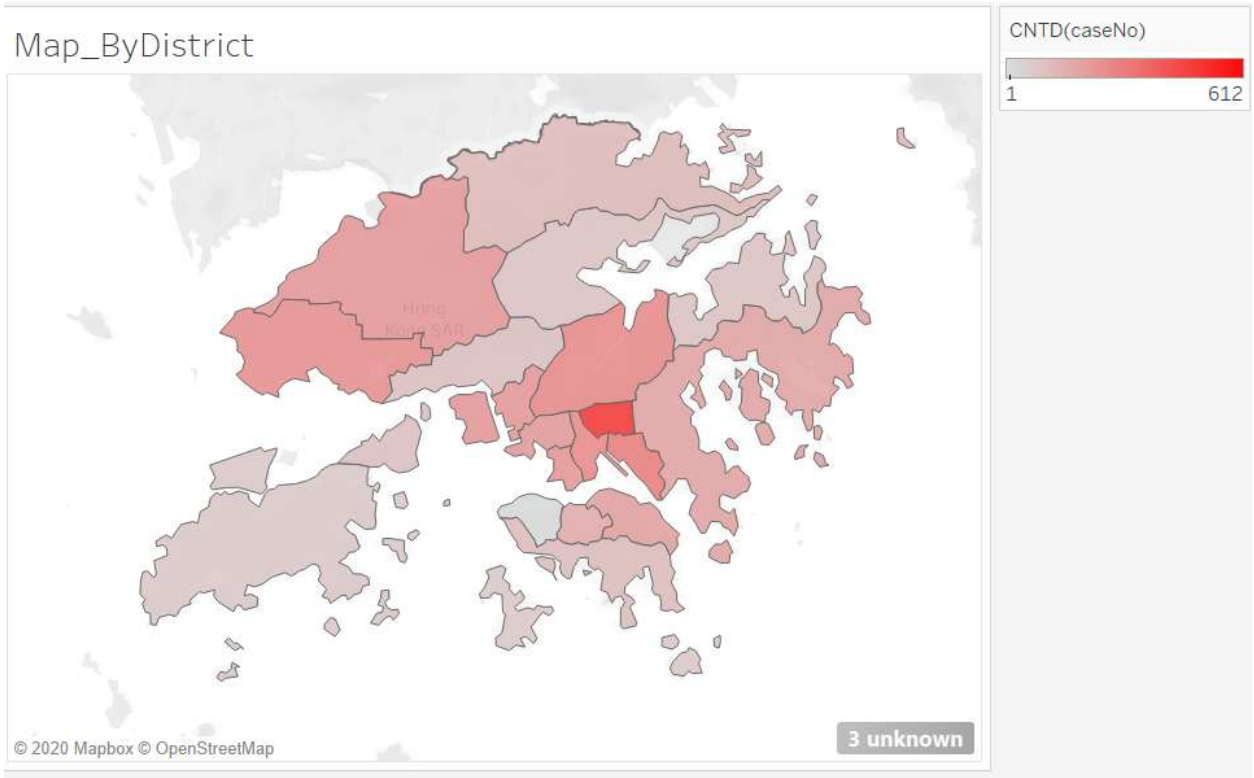
Mandatory task

Visualize the following topics

1. Time series analysis on number of onset case

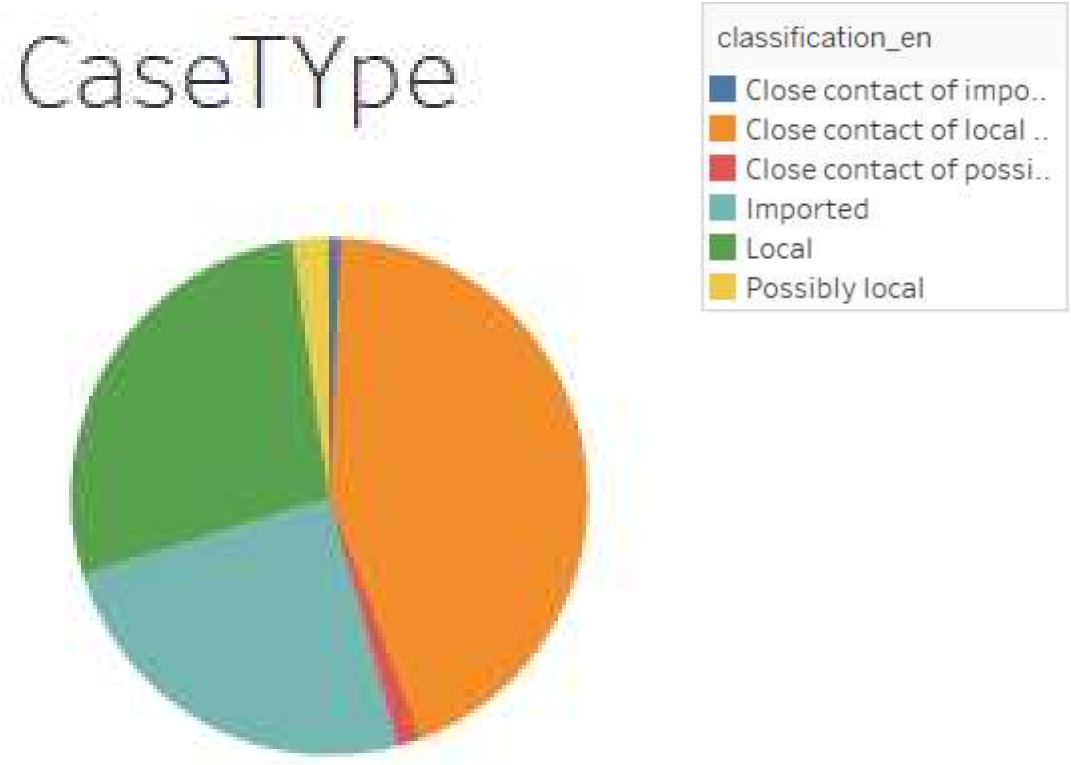


2. Number of cases by district

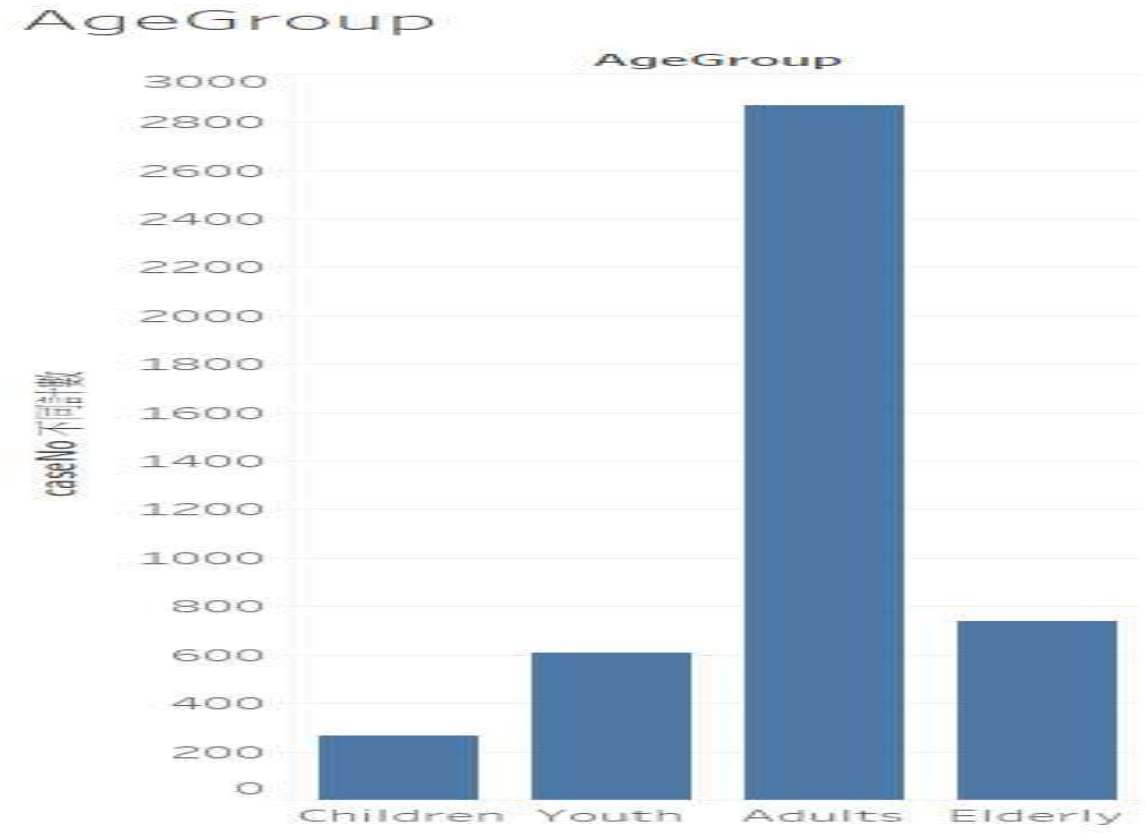


3. Keyword Cloud of Reported case

4. Contribution of case type



5. Count of age group[1]



Age Category	Range
Children	Younger than 14
Youth	15 - 24
Adults	25-64
Elderly	Older than 64

Advanced technique task

This task is freestyle from your idea. You must create one of dashboard is consolidate some of visualize graph from last section and drill some valuable findings. Furthermore, you still have to plot some graphs that not mentioned from last section and consolidate them as story-telling dashboard.

[1] Age group reference: https://www.statcan.gc.ca/eng/concepts/definitions/age2

Finding demonstrations.

You should present your idea/findings through story mode in 20 mins. No PowerPoint or other presentation software allowed. The presentation should make use of following Tableau features:

1. Story mode
2. Dashboard
3. Filtering data
4. Drill-in data
5. Story telling
6. Presentation Mode
7. Single Graph sheet

Furthermore, you also have prepared a report for our record. The template is attached in Moodle. The Report have to include the following parts:

1. Introduction of project
2. Data Contents
3. Findings
4. Suggestions
5. Reference (if any)

Project Case 2122 Full Document

You are required to form a group of 2~ 3 classmates or even work individually. Each group or individual will receive one set of data. All groups or individuals have to process the dataset and generate some findings. The data is about the transaction data of British Property. All groups and individuals use the provided data and finish the following tasks: Data preparation, Mandatory task, Advanced technique task, and Finding demonstrations. All tasks and demonstrations are required to utilize the features of Tableau Public and R/Python for data pre-processing.

To Conduct the project smoothly, you may have some questions for your starting point, and the following question is only an **example:**

1. Which type of property is popular in the UK?
2. How is the pandemic affecting property transactions?
3. Which kind of property will become the next trend?

Dataset Description

Transaction record from the Government

This is public data from the UK Government Open Data website[2]. The data contain the details of each transaction, from 1995 to 2021. The following is a sample of the dataset.

T id	Price	Dateof Trans	Postalcode	Type	Old New	Duration	Paon	Saon	Street	Locality	Town City	District
{A2479555-4F39-74C...	250,000	7/2/2020 0:00:00	EN8 8SA	Maisonettes	Established residential	Leasehold	EDWARDS COURT	FLAT 18	TURNERS HILL	null	WALTHAM CROSS	BROXBOURNE
{A2479555-4F3A-74C...	370,000	14/2/2020 0:00:00	WD6 1UG	Terraced	Established residential	Freehold	4	null	WELBECK CLOSE	null	BOREHAMWOOD	HERTSMERE
{A2479555-4F3B-74C...	685,000	6/3/2020 0:00:00	AL6 0PW	Detached	Established residential	Freehold	13	null	THE AVENUE	null	WELWYN	WELWYN HATFIELD
{A2479555-4F3C-74C...	133,000	10/1/2020 0:00:00	SG1 4EL	Maisonettes	Established residential	Leasehold	384	null	YORK ROAD	null	STEVENAGE	STEVENAGE
{A2479555-4F3D-74C...	223,000	28/2/2020 0:00:00	SG2 7QR	Terraced	Established residential	Freehold	14	null	CONIFER WALK	null	STEVENAGE	STEVENAGE
{A2479555-4F3E-74C...	725,000	21/2/2020 0:00:00	SG13 7TQ	Detached	Established residential	Freehold	2	null	HONEYSUCKLE CLOSE	null	HERTFORD	EAST HERTFORDSHIF
{A2479555-4F3F-74C...	2,075,000	28/2/2020 0:00:00	AL1 4BG	Detached	Established residential	Freehold	23	null	HOMEWOOD ROAD	null	ST ALBANS	ST ALBANS
{A2479555-4F40-74C...	708,200	25/2/2020 0:00:00	CM23 5TB	Detached	Established residential	Freehold	24	null	WRAGLINGS	null	BISHOP'S STORTFORD	EAST HERTFORDSHIF
{A2479555-4F41-74C...	258,000	28/2/2020 0:00:00	WD6 3HL	Maisonettes	Established residential	Leasehold	VALENTINE COURT	3	BARNET LANE	ELSTREE	BOREHAMWOOD	HERTSMERE
{A2479555-4F42-74C...	270,000	20/2/2020 0:00:00	HP1 2TW	Terraced	Established residential	Freehold	78	null	THE PASTURES	null	HEMEL HEMPSTEAD	DACORUM
{A2479555-4F43-74C...	557,500	21/2/2020 0:00:00	SG5 1UB	Terraced	Established residential	Freehold	139	null	BEARTON ROAD	null	HITCHIN	NORTH HERTFORDS.
{A2479555-4F44-74C...	295,000	20/2/2020 0:00:00	HP2 6LL	Terraced	Established residential	Freehold	28	null	CLAYMORE	null	HEMEL HEMPSTEAD	DACORUM
{A2479555-4F45-74C...	319,995	28/2/2020 0:00:00	EN8 8TS	Terraced	Established residential	Freehold	11	null	RUSSELLS RIDE	CHESHUNT	WALTHAM CROSS	BROXBOURNE
{A2479555-4F46-74C...	265,500	10/2/2020 0:00:00	CM23 2AJ	Terraced	Established residential	Freehold	7	null	CHERRY GARDENS	null	BISHOP'S STORTFORD	EAST HERTFORDSHIF

[2] Data source website: https://www.gov.uk/government/statistical-data-sets/price-paid-data-downloads#single-file

Data Description

The following is the Data Description from the UK Open data website, and this information may help you summarise the data dictionary for this project.

Data item	Explanation (where appropriate)
Transaction unique identifier	A reference number which is generated automatically recording each published sale. The number is unique and will change each time a sale is recorded.
Price	Sale price stated on the transfer deed.
Date of Transfer	Date when the sale was completed, as stated on the transfer deed.
Postcode	This is the postcode used at the time of the original transaction. Note that postcodes can be reallocated and these changes are not reflected in the Price Paid Dataset.
Property Type	D = Detached, S = Semi-Detached, T = Terraced, F = Flats/Maisonettes, O = Other Note that: - we only record the above categories to describe property type, we do not separately identify bungalows - end-of-terrace properties are included in the Terraced category above - 'Other' is only valid where the transaction relates to a property type that is not covered by existing values, for example where a property comprises more than one large parcel of land
Old/New	Indicates the age of the property and applies to all price paid transactions, residential and non-residential. Y = a newly built property, N = an established residential building

Data item	Explanation (where appropriate)
Duration	Relates to the tenure: F = Freehold, L= Leasehold etc. Note that HM Land Registry does not record leases of 7 years or less in the Price Paid Dataset.
PAON	Primary Addressable Object Name. Typically the house number or name.
SAON	Secondary Addressable Object Name. Where a property has been divided into separate units (for example, flats), the PAON (above) will identify the building and a SAON will be specified that identifies the separate unit/flat.
Street	
Locality	
Town/City	
District	
County	
PPD Category Type	Indicates the type of Price Paid transaction. A = Standard Price Paid entry, includes single residential property sold for value. B = Additional Price Paid entry including transfers under a power of sale/repossessions, buy-to-lets (where they can be identified by a Mortgage), transfers to non-private individuals and sales where the property type is classed as 'Other'. Note that category B does not separately identify the transaction types

Data item	Explanation (where appropriate)
	stated. HM Land Registry has been collecting information on Category A transactions from January 1995. Category B transactions were identified from October 2013.
Record Status - monthly file only	Indicates additions, changes and deletions to the records.(see guide below). A = Addition C = Change D = Delete Note that where a transaction changes category type due to misallocation (as above) it will be deleted from the original category type and added to the correct category with a new transaction unique identifier.

Task

Data preparation

You will receive 26,321,785 transaction records for your project, and you have to ETL before pouring the data into Tableau. Therefore, create a R/Python notebook to:

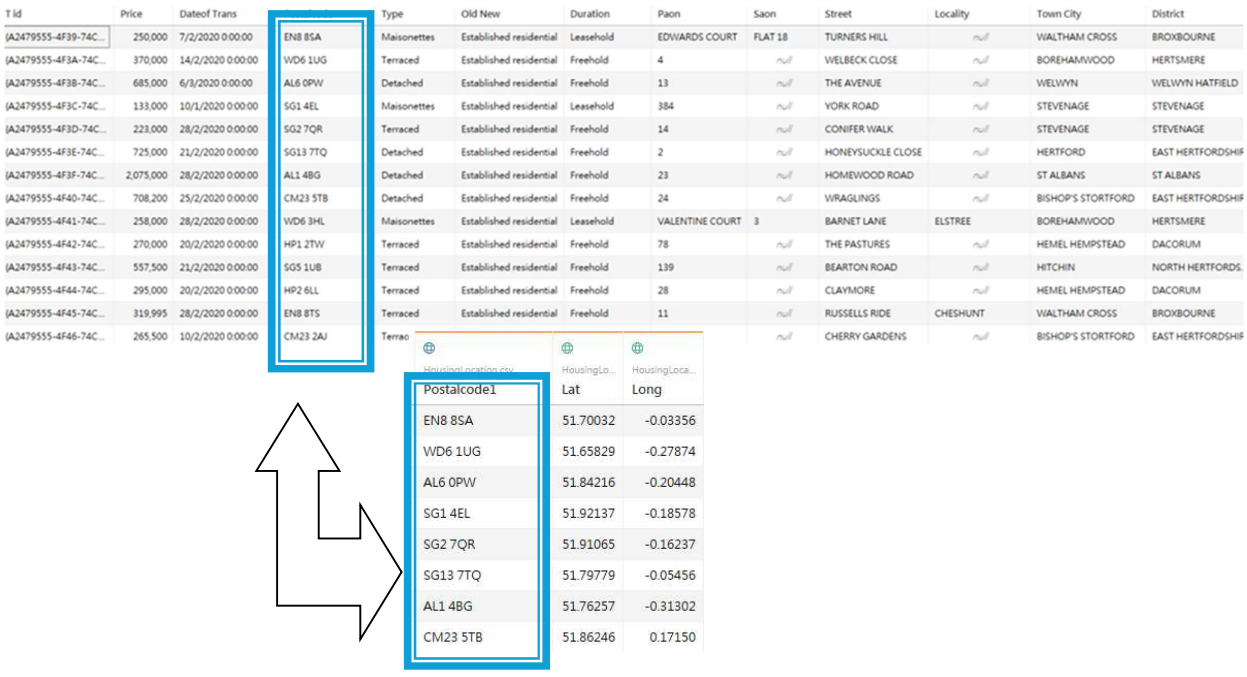
3. Extract the data during the pandemic (The year 2020 to the Year 2021).
4. Since the dataset does not contain headers, you must label all headers based on the information provided.
5. You have to convert all Category codes to Category labels.
6. Save your result to a CSV format file named Housing_Final.csv
7. Find all geographical data for all properties as much as possible and save as a single dataset called HousingLocation.csv

Deliverable:

1. Several dataset files in a CSV format:
   I. Main Data should be in one file (Housing_Final.csv)
   II. Supplementary data should separate data files (HousingLocation.csv)

Hint:

3. Filtrate all unnecessary columns (up to you)
4. Postal Codes can help you to find all property geographical data.
5. The below is just for your reference (in a CSV format)



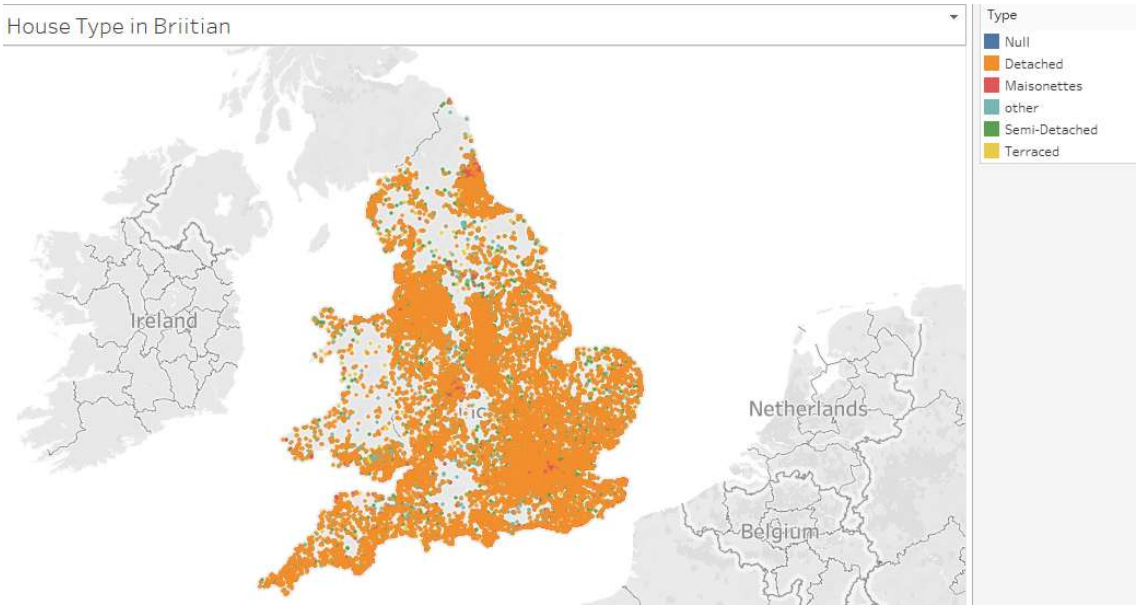
T id	Price	Dateof Trans		Type	Old New	Duration	Paon	Saon	Street	Locality	Town City	District
{A2479555-4F39-74C...	250,000	7/2/2020 0:00:00	EN8 8SA	Maisonettes	Established residential	Leasehold	EDWARDS COURT	FLAT 18	TURNERS HILL		WALTHAM CROSS	BROXBOURNE
{A2479555-4F3A-74C...	370,000	14/2/2020 0:00:00	WD6 1UG	Terraced	Established residential	Freehold	4		WELBECK CLOSE		BOREHAMWOOD	HERTSMERE
{A2479555-4F3B-74C...	685,000	6/3/2020 0:00:00	AL6 0PW	Detached	Established residential	Freehold	13		THE AVENUE		WELWYN	WELWYN HATFIELD
{A2479555-4F3C-74C...	133,000	10/1/2020 0:00:00	SG1 4EL	Maisonettes	Established residential	Leasehold	384		YORK ROAD		STEVENAGE	STEVENAGE
{A2479555-4F3D-74C...	223,000	28/2/2020 0:00:00	SG2 7QR	Terraced	Established residential	Freehold	14		CONIFER WALK		STEVENAGE	STEVENAGE
{A2479555-4F3E-74C...	725,000	21/2/2020 0:00:00	SG13 7TQ	Detached	Established residential	Freehold	2		HONEYSUCKLE CLOSE		HERTFORD	EAST HERTFORDSHIF
{A2479555-4F3F-74C...	2,075,000	28/2/2020 0:00:00	AL1 4BG	Detached	Established residential	Freehold	23		HOMEWOOD ROAD		ST ALBANS	ST ALBANS
{A2479555-4F40-74C...	708,200	25/2/2020 0:00:00	CM23 5TB	Detached	Established residential	Freehold	24		WRAGLINGS		BISHOP'S STORTFORD	EAST HERTFORDSHIF
{A2479555-4F41-74C...	258,000	28/2/2020 0:00:00	WD6 3HL	Maisonettes	Established residential	Leasehold	VALENTINE COURT	3	BARNET LANE	ELSTREE	BOREHAMWOOD	HERTSMERE
{A2479555-4F42-74C...	270,000	20/2/2020 0:00:00	HP1 2TW	Terraced	Established residential	Freehold	78		THE PASTURES		HEMEL HEMPSTEAD	DACORUM
{A2479555-4F43-74C...	557,500	21/2/2020 0:00:00	SG5 1UB	Terraced	Established residential	Freehold	139		BEARTON ROAD		HITCHIN	NORTH HERTFORDS.
{A2479555-4F44-74C...	295,000	20/2/2020 0:00:00	HP2 6LL	Terraced	Established residential	Freehold	28		CLAYMORE		HEMEL HEMPSTEAD	DACORUM
{A2479555-4F45-74C...	319,995	28/2/2020 0:00:00	EN8 8TS	Terraced	Established residential	Freehold	11		RUSSELLS RIDE	CHESHUNT	WALTHAM CROSS	BROXBOURNE
{A2479555-4F46-74C...	265,500	10/2/2020 0:00:00	CM23 2AJ	Terrac					CHERRY GARDENS		BISHOP'S STORTFORD	EAST HERTFORDSHIF

Postalcode1	Lat	Long
EN8 8SA	51.70032	-0.03356
WD6 1UG	51.65829	-0.27874
AL6 0PW	51.84216	-0.20448
SG1 4EL	51.92137	-0.18578
SG2 7QR	51.91065	-0.16237
SG13 7TQ	51.79779	-0.05456
AL1 4BG	51.76257	-0.31302
CM23 5TB	51.86246	0.17150

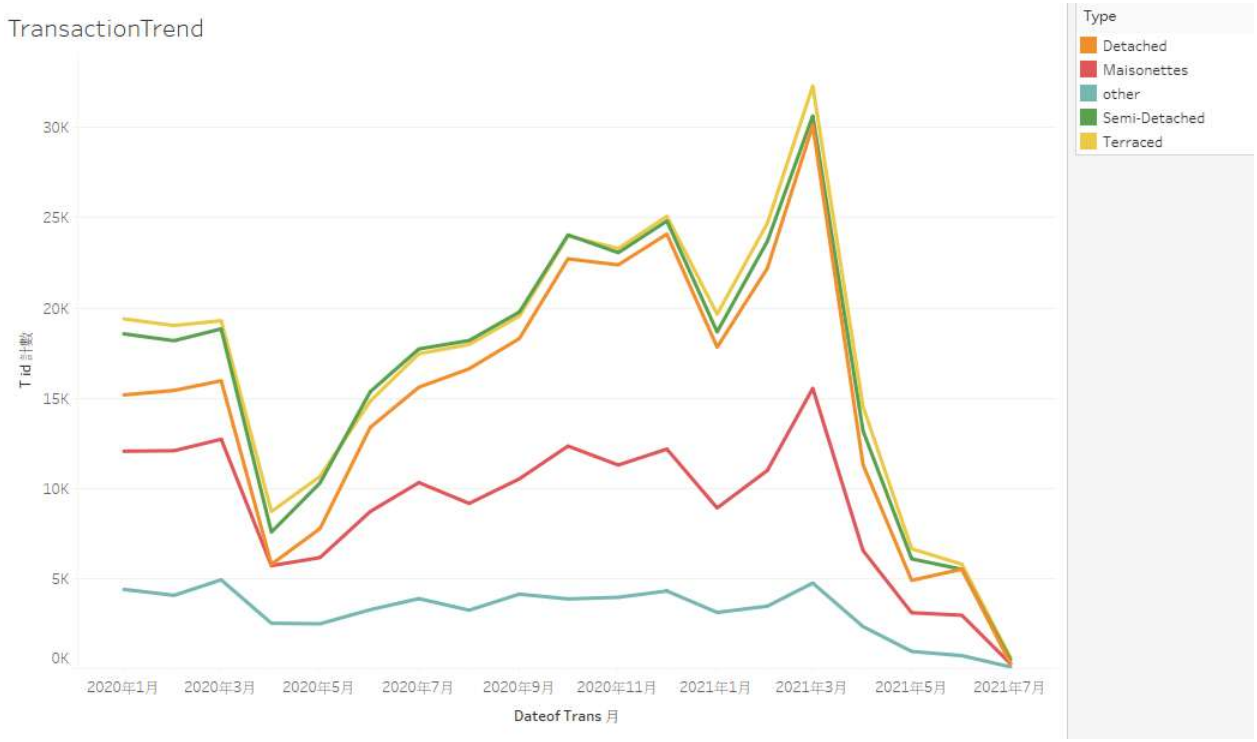
Mandatory task

Visualize the following requirements

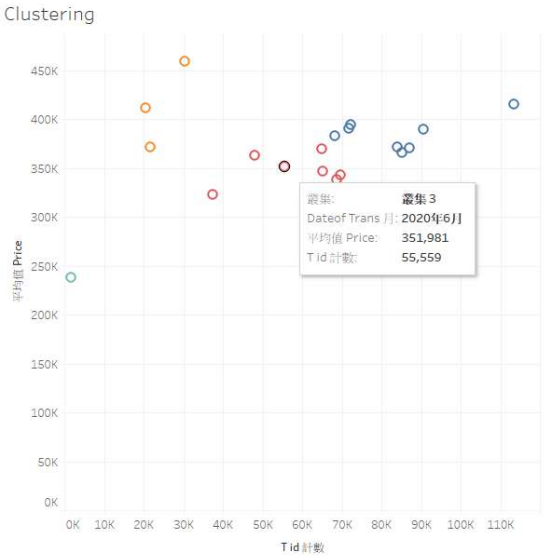
6. The property distribution in the UK.



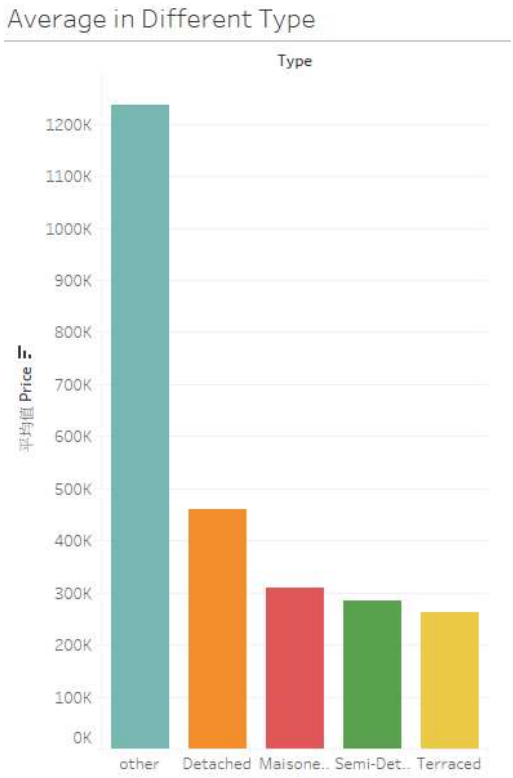
7. The trend of the number of the transaction by type of house.



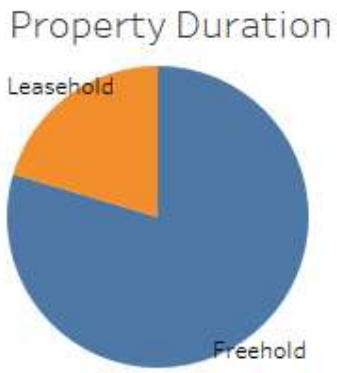
8. Number of Transaction VS Average Price of Transaction (The Datapoint is transaction period on a monthly basis)



9. Average Price of each Type of Property

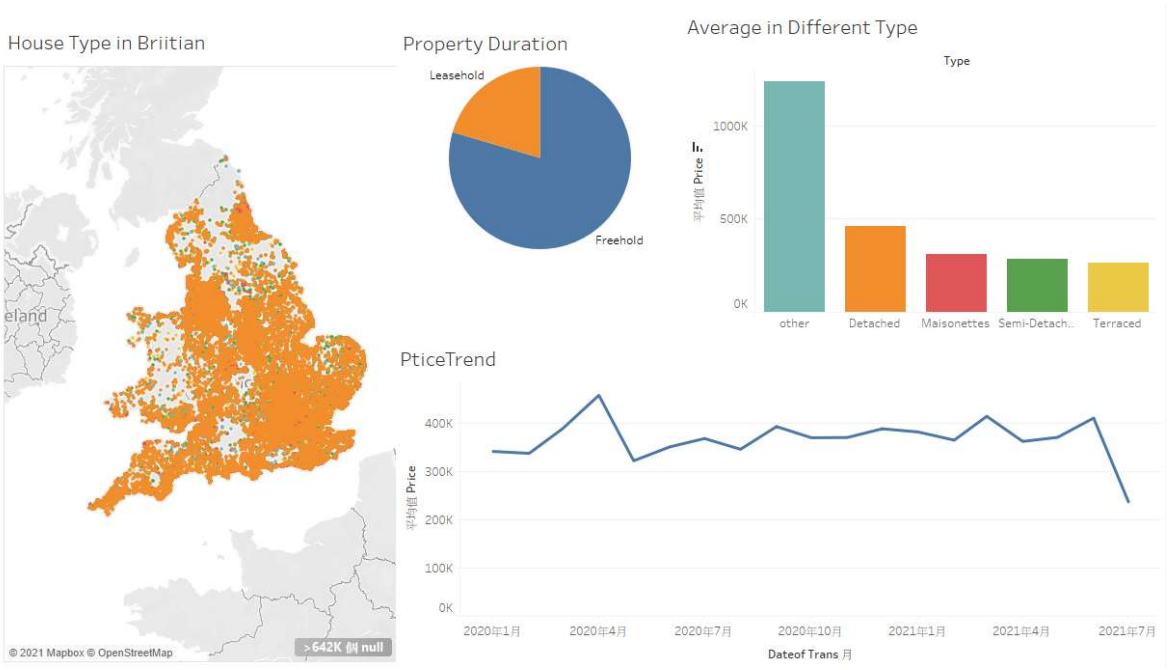


10. The proportion of Property Duration types.



Advanced technique task

This task is a freestyle from your idea. First, you have to create dashboards for your project purpose. Bear in mind that you must complete one of the dashboards by using visualized plottings from the last section and drilling some valuable findings. Furthermore, you still have to plot some graphs not mentioned in the previous section and consolidate them as a story-telling dashboard. The below is a dashboard for your reference. You may follow the guidelines of Steve Few informative dashboards for designing a good dashboard. Remember, a good dashboard is not an ornament; it should be a tool for assisting your findings. PLEASE DON'T USE A SINGLE TYPE OF GRAPH FOR YOUR DASHBOARD.



Finding demonstrations

You should present your idea/findings through the story mode in 15 mins. No PowerPoint or other presentation software is allowed. The presentation should make use of the following Tableau features:

1. Story mode
2. Dashboard
3. Filtering data
4. Drill-in data
5. Story-telling
6. Presentation Mode
7. Single Graph sheet

Furthermore, you also have prepared a report for our record. You may find the report template in Moodle. The Report have to include the following parts:

1. Introduction of project
2. Data Dictionary
3. Findings
4. Suggestions
5. Reference (if any)

Project Overview

With the development of information technology, people's recreation is highly reliant on social media platforms. A huge amount of user behaviour records are generated on the platforms. Therefore, understanding the platform user's behaviour can effectively promote your product to the target customers/audience.

You are required to form a group of 2~ 3 classmates or even work individually. Each group or individual will receive one set of data. All groups or individuals have to process the dataset and generate some findings. The data is about YouTube video data from various countries. All groups and individuals use the provided data.

To Conduct the project smoothly, you may have some questions for your starting point, and the following question is only an **example:**

1. Comparison between two regions of the video?
2. What is the audience interested in?
3. What is YouTube used for?
4. Is virtual YouTube popular in the Western world?

Dataset Description

Youtube video record

This is a open data from data organisation Kaggele.com. The data contain the details of each uploaded video from 2020 to 2022. The following is a sample of the dataset.

video_id	title	publishedAt	channelId	channelTitl	trending_d	tags	view_coun	likes	dislikes	comment_	comments_	ratings_dis
yVdH3Qac	Selena Gor	10/8/2020 16:32	UCPNxhD'	Selena Gor	12/8/2020	Selena Gor	1523818	163686	2377	9845	FALSE	FALSE
S5PotZWS	Pi?lula da	6/8/2020 20:11	UCqppqPJl	Luarices	12/8/2020	luaricesllua	817468	121100	850	6794	FALSE	FALSE
yVdH3Qac	Selena Gor	10/8/2020 16:32	UCPNxhD'	Selena Gor	13/8/2020	Selena Gor	1738170	183806	2552	10352	FALSE	FALSE
yVdH3Qac	Selena Gor	10/8/2020 16:32	UCPNxhD'	Selena Gor	14/8/2020	Selena Gor	1871931	193333	2655	10609	FALSE	FALSE
yVdH3Qac	Selena Gor	10/8/2020 16:32	UCPNxhD'	Selena Gor	15/8/2020	Selena Gor	1991246	202347	2729	10782	FALSE	FALSE
yVdH3Qac	Selena Gor	10/8/2020 16:32	UCPNxhD'	Selena Gor	16/8/2020	Selena Gor	2081065	208991	2764	10946	FALSE	FALSE
fbS0efVvk	Batman I T	23/8/2020 1:01	UC5XG4y	Ingresso.cc	23/8/2020	Ingresso.cc	537407	64800	705	5481	FALSE	FALSE
xXEwp0uC	Mulher-Ma	22/8/2020 17:25	UCEOVI4	Warner Bro	23/8/2020	Mulher-Ma	317606	19338	199	795	FALSE	FALSE
x7cYkXvz	O ESQUAI	22/8/2020 19:24	UCEOVI4	Warner Bro	23/8/2020	The Suicid	178504	12602	133	565	FALSE	FALSE
fbS0efVvk	Batman I T	23/8/2020 1:01	UC5XG4y	Ingresso.cc	24/8/2020	Ingresso.cc	1448014	100475	2092	7999	FALSE	FALSE
xXEwp0uC	Mulher-Ma	22/8/2020 17:25	UCEOVI4	Warner Bro	24/8/2020	Mulher-Ma	509235	24673	327	993	FALSE	FALSE
x7cYkXvz	O ESQUAI	22/8/2020 19:24	UCEOVI4	Warner Bro	24/8/2020	The Suicid	324999	16621	247	731	FALSE	FALSE
fbS0efVvk	Batman I T	23/8/2020 1:01	UC5XG4y	Ingresso.cc	25/8/2020	Ingresso.cc	1830088	114688	2363	9064	FALSE	FALSE
xXEwp0uC	Mulher-Ma	22/8/2020 17:25	UCEOVI4	Warner Bro	25/8/2020	Mulher-Ma	547752	25299	334	1042	FALSE	FALSE

Data Description

The following is the Data Description from the Open Data website, and this information may help you summarise the data dictionary for this project.

Data item	Explanation (where appropriate)
video_id	Unique Code of the video when uploaded to / created by Youtube.
title	Title of the video
publishedAt	The date that the video published
channelId	The video that channel belongs to.
channelTitle	The name of the channel
trending_date	The date that video becomes hot worldwide

Data item	Explanation (where appropriate)
view_count	The number of view
likes	The number of likes
dislikes	The number of dislike
comment_count	The number of comment
comments_disabled	Is the video disabled the comment function?
ratings_disabled	Is the video disable the rating feature?
IDdescription	Description of the video
tag	Tags of the videos

Task

Data Preparation

You will receive 11 countries of YouTube video data for your project, you have to ETL before pouring the data into Tableau:

Stage 1

1. You may merge the different countries of data, merged content is based on your project purpose but at least two countries. (E.g. : Great British and United States)
2. Convert the Category ID to Category Name, the Category Name included in the XX_category_id.json[3] file

Before

After

[3] All json file contains identical content.

3. Label the region of the video source, the below is the table of the label. (Hints: the region of the video source is based on the data file name. like BR_youtube_trending_data.csv is from Brazil.

Label	Description
BR	Brazil
CA	Canada
DE	Demark
FR	France
GB	Great British
IN	India
JP	Japan
KR	Korea
MX	Mexico
RU	Russia
US	United State

Stage 2

1. Split the tags to the individual data row.

Before

ps5|playstation 5|playstation 5 digital edition|ps5 digital edition|playstation|ps5 digital|playstation 5 digital|ps5 price|ps5 release|sony|ps5 vs xbox|ps5 vs series x|ps5 vs series s|ps5 announcement|ps5 event|playstation 5 vs xbox|xbox|microsoft|console|game console|the test drivers|test drivers|this is|austin evans

After

```
"ps5"
"playstation 5"
"playstation 5 digital edition"
"ps5 digital edition"
"playstation"
"ps5 digital"
"playstation 5 digital"
"ps5 price"
"ps5 release"
"sony"
"ps5 vs xbox"
"ps5 vs series x"
"ps5 vs series s"
"ps5 announcement"
"ps5 event"
"playstation 5 vs xbox"
"xbox"
"microsoft"
"console"
"game console"
"the test drivers"
"test drivers"
"this is"
"austin evans"
```

Hint:

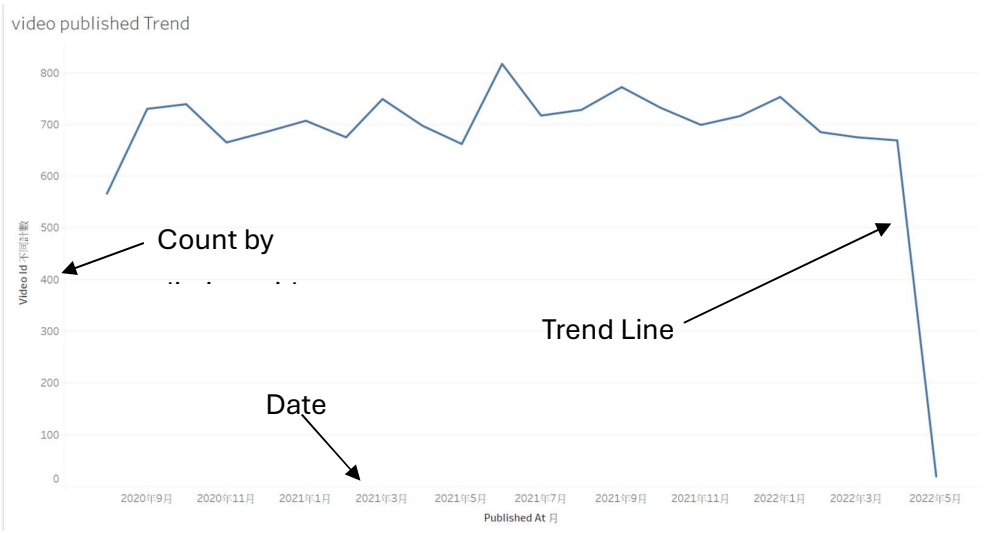
1. Filtrate all unnecessary keyword and select appropriate columns (up to you)
2. Use fread() to read the csv files, **PLEASE DON'T USE READ.CSV at this moment.**
3. **Due to the dataset is too large, please select appropriate data file for your project and split the data to your teammate for preparation and combine them when finished.**
   i. **My notebook can process 20000 video tags per 30 minutes.**
4. The below data sample just for your reference only.


```
"","video_id","title","publishedAt","channelId","channelTitle","trending_date","view_count","likes","dislikes","comment_count","comments_disabled","ratings_disabled","Region","IDdescription","tagSplit
"1","QXeZUeD0Xo4","□□□ Han Seung Woo Sacrifice MV",2020-08-10 09:00:00,"UCQ7a83KJBE7w_txsT7VaJ8g","VICTON 빅톤",2020-08-13,4877120,124842,496,9740,FALSE,FALSE,"England","Entertainment","빅톤"
"2","QXeZUeD0Xo4","□□□ Han Seung Woo Sacrifice MV",2020-08-10 09:00:00,"UCQ7a83KJBE7w_txsT7VaJ8g","VICTON 빅톤",2020-08-13,4877120,124842,496,9740,FALSE,FALSE,"England","Entertainment","한승우"
"3","QXeZUeD0Xo4","□□□ Han Seung Woo Sacrifice MV",2020-08-10 09:00:00,"UCQ7a83KJBE7w_txsT7VaJ8g","VICTON 빅톤",2020-08-13,4877120,124842,496,9740,FALSE,FALSE,"England","Entertainment","승우"
"4","QXeZUeD0Xo4","□□□ Han Seung Woo Sacrifice MV",2020-08-10 09:00:00,"UCQ7a83KJBE7w_txsT7VaJ8g","VICTON 빅톤",2020-08-13,4877120,124842,496,9740,FALSE,FALSE,"England","Entertainment","하찬"
"5","QXeZUeD0Xo4","□□□ Han Seung Woo Sacrifice MV",2020-08-10 09:00:00,"UCQ7a83KJBE7w_txsT7VaJ8g","VICTON 빅톤",2020-08-13,4877120,124842,496,9740,FALSE,FALSE,"England","Entertainment","임세준"
"6","QXeZUeD0Xo4","□□□ Han Seung Woo Sacrifice MV",2020-08-10 09:00:00,"UCQ7a83KJBE7w_txsT7VaJ8g","VICTON 빅톤",2020-08-13,4877120,124842,496,9740,FALSE,FALSE,"England","Entertainment","도한세"
"7","QXeZUeD0Xo4","□□□ Han Seung Woo Sacrifice MV",2020-08-10 09:00:00,"UCQ7a83KJBE7w_txsT7VaJ8g","VICTON 빅톤",2020-08-13,4877120,124842,496,9740,FALSE,FALSE,"England","Entertainment","최병찬"
"8","QXeZUeD0Xo4","□□□ Han Seung Woo Sacrifice MV",2020-08-10 09:00:00,"UCQ7a83KJBE7w_txsT7VaJ8g","VICTON 빅톤",2020-08-13,4877120,124842,496,9740,FALSE,FALSE,"England","Entertainment","정수빈"
"9","QXeZUeD0Xo4","□□□ Han Seung Woo Sacrifice MV",2020-08-10 09:00:00,"UCQ7a83KJBE7w_txsT7VaJ8g","VICTON 빅톤",2020-08-13,4877120,124842,496,9740,FALSE,FALSE,"England","Entertainment","hanseungwoo"
"10","QXeZUeD0Xo4","□□□ Han Seung Woo Sacrifice MV",2020-08-10 09:00:00,"UCQ7a83KJBE7w_txsT7VaJ8g","VICTON 빅톤",2020-08-13,4877120,124842,496,9740,FALSE,FALSE,"England","Entertainment","kangseungsik"
"11","QXeZUeD0Xo4","□□□ Han Seung Woo Sacrifice MV",2020-08-10 09:00:00,"UCQ7a83KJBE7w_txsT7VaJ8g","VICTON 빅톤",2020-08-13,4877120,124842,496,9740,FALSE,FALSE,"England","Entertainment","heochan"
"12","QXeZUeD0Xo4","□□□ Han Seung Woo Sacrifice MV",2020-08-10 09:00:00,"UCQ7a83KJBE7w_txsT7VaJ8g","VICTON 빅톤",2020-08-13,4877120,124842,496,9740,FALSE,FALSE,"England","Entertainment","limsejun"
"13","QXeZUeD0Xo4","□□□ Han Seung Woo Sacrifice MV",2020-08-10 09:00:00,"UCQ7a83KJBE7w_txsT7VaJ8g","VICTON 빅톤",2020-08-13,4877120,124842,496,9740,FALSE,FALSE,"England","Entertainment","dohanse"
"14","QXeZUeD0Xo4","□□□ Han Seung Woo Sacrifice MV",2020-08-10 09:00:00,"UCQ7a83KJBE7w_txsT7VaJ8g","VICTON 빅톤",2020-08-13,4877120,124842,496,9740,FALSE,FALSE,"England","Entertainment","choibyungchan"
"15","QXeZUeD0Xo4","□□□ Han Seung Woo Sacrifice MV",2020-08-10 09:00:00,"UCQ7a83KJBE7w_txsT7VaJ8g","VICTON 빅톤",2020-08-13,4877120,124842,496,9740,FALSE,FALSE,"England","Entertainment","jungsubin"
"16","QXeZUeD0Xo4","□□□ Han Seung Woo Sacrifice MV",2020-08-10 09:00:00,"UCQ7a83KJBE7w_txsT7VaJ8g","VICTON 빅톤",2020-08-13,4877120,124842,496,9740,FALSE,FALSE,"England","Entertainment","컨티뉴어스"
"17","QXeZUeD0Xo4","□□□ Han Seung Woo Sacrifice MV",2020-08-10 09:00:00,"UCQ7a83KJBE7w_txsT7VaJ8g","VICTON 빅톤",2020-08-13,4877120,124842,496,9740,FALSE,FALSE,"England","Entertainment","continuous"
"18","QXeZUeD0Xo4","□□□ Han Seung Woo Sacrifice MV",2020-08-10 09:00:00,"UCQ7a83KJBE7w_txsT7VaJ8g","VICTON 빅톤",2020-08-13,4877120,124842,496,9740,FALSE,FALSE,"England","Entertainment","하울링"
"19","QXeZUeD0Xo4","□□□ Han Seung Woo Sacrifice MV",2020-08-10 09:00:00,"UCQ7a83KJBE7w_txsT7VaJ8g","VICTON 빅톤",2020-08-13,4877120,124842,496,9740,FALSE,FALSE,"England","Entertainment","howling"
"20","QXeZUeD0Xo4","□□□ Han Seung Woo Sacrifice MV",2020-08-10 09:00:00,"UCQ7a83KJBE7w_txsT7VaJ8g","VICTON 빅톤",2020-08-13,4877120,124842,496,9740,FALSE,FALSE,"England","Entertainment","Mayday"
"21","QXeZUeD0Xo4","□□□ Han Seung Woo Sacrifice MV",2020-08-10 09:00:00,"UCQ7a83KJBE7w_txsT7VaJ8g","VICTON 빅톤",2020-08-13,4877120,124842,496,9740,FALSE,FALSE,"England","Entertainment","메이데이"
"22","QXeZUeD0Xo4","□□□ Han Seung Woo Sacrifice MV",2020-08-10 09:00:00,"UCQ7a83KJBE7w_txsT7VaJ8g","VICTON 빅톤",2020-08-13,4877120,124842,496,9740,FALSE,FALSE,"England","Entertainment","Sacrifice"
"23","QXeZUeD0Xo4","□□□ Han Seung Woo Sacrifice MV",2020-08-10 09:00:00,"UCQ7a83KJBE7w_txsT7VaJ8g","VICTON 빅톤",2020-08-13,4877120,124842,496,9740,FALSE,FALSE,"England","Entertainment","Fame"
```

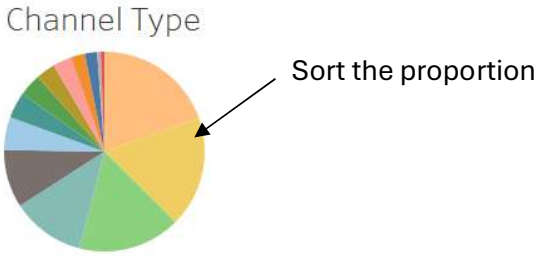

Mandatory task (4% Ea.)

Visualize the following requirements, the below plotting only for your reference.

1. The time series trend line on number of video published



2. Proportion of Channel Type (Category).

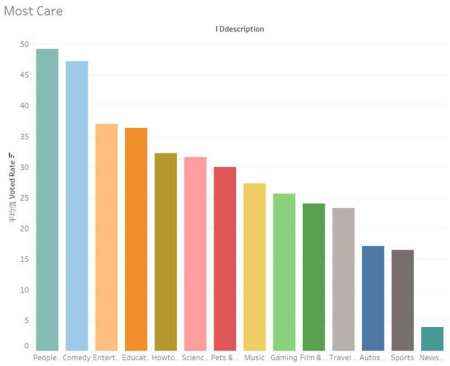


3. Tags Keyword Cloud

Font size of the keyword tags is depending on the frequency of the

The range of frequency is from 100 to 1000

4. Average voted Rate



Hints

Equations

#Vote Rate is a real response to have liked or disliked the video after watching it.

Vote Rate = Replied / View Count

#Replied is the population of users who like or dislike your video.

Replied = Likes + Dislike

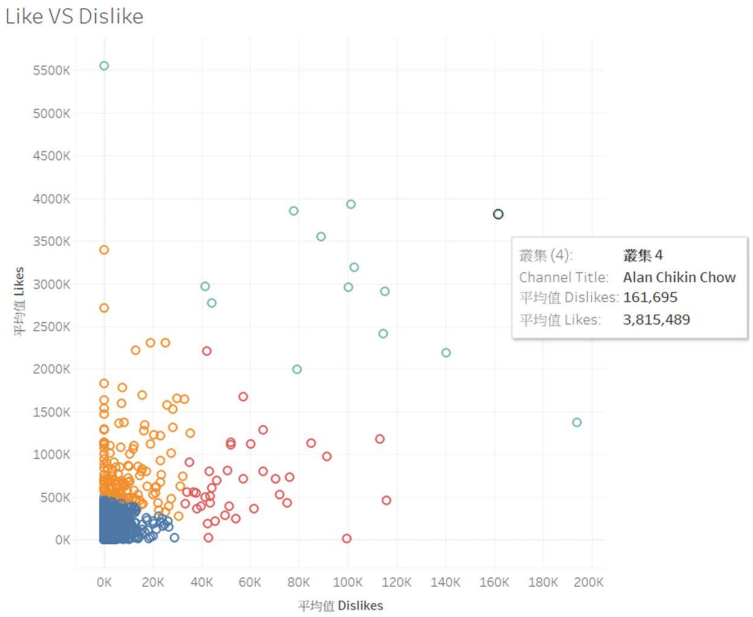
How to fixed if the value is too large?

There are many duplicated IDs after splitting the tags, and you may use FIXED to tell the calculated field only needs the first data of the duplicated Dimension for manipulation.

Syntax : {mixed [Dimension]: min([measures])}

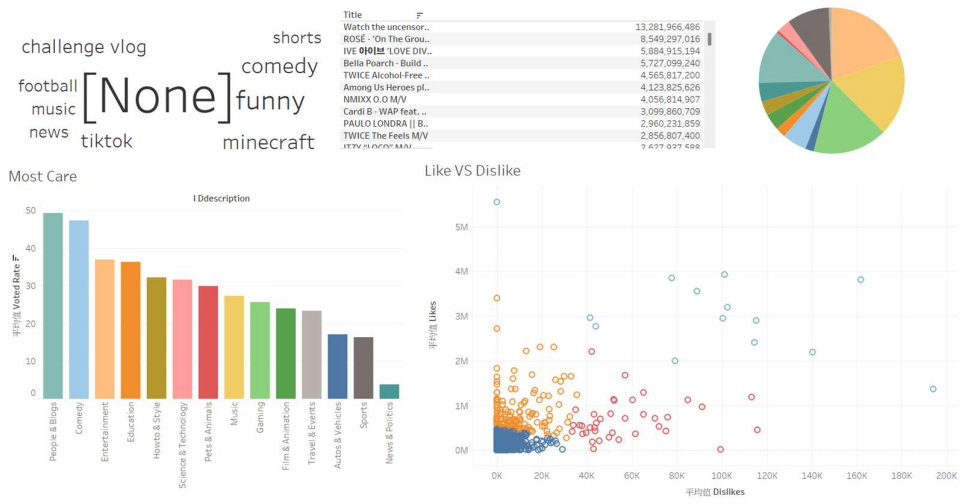
Dimension	measures
A	1
A	1
B	2
C	2
C	3

5. Clustering of the relationship between average Likes and average Dislike by Channel



Advanced technique task

This task is a freestyle from your idea. First, you have to create dashboards for your project purpose. Remember that you must complete one of the dashboards by using visualised plottings from the last section and drilling some valuable findings. Furthermore, you still have to plot some graphs not mentioned in the previous section and consolidate them as a story-telling dashboard. Below is a dashboard for your reference. You may follow the guideline of Steve Few's informative dashboards for designing a good dashboard. Remember, a good dashboard is not an ornament; it should be a tool for assisting your findings. PLEASE DON'T USE A SINGLE TYPE OF GRAPH FOR YOUR DASHBOARD.



Finding demonstrations

You should present your idea/findings through the story mode in 15 mins. No PowerPoint or other presentation software is allowed. The presentation should make use of the following Tableau features:

1. Story mode
2. Dashboard
3. Filtering data
4. Drill-in data
5. Story-telling
6. Presentation Mode
7. Single Graph sheet

Furthermore, you also have prepared a report for our record. You may find the report template in Moodle. The Report has to include the following parts:

1. Introduction of project
2. Data Dictionary
3. Findings
4. Suggestions
5. Reference (if any)

Project Case 2324 Full Document

Project Overview

You are required to form a group of 2~ 3 classmates or even work individually. Each group or individual will receive one set of data. All groups or individuals have to process the dataset and generate some findings. The data is about LA Crime Data in the past 13 years. All groups and individuals use the provided data.

To Conduct the project smoothly, you may have some questions for your starting point, and the following question is only an **example:**

1. What are the Crime Change in these 10 Years?
2. A Safety Instruction for tourists?
3. Which area is avoid to visit?
4. What are the circumstances of crime in LA?

Dataset Description

Youtube video record

This is open data from the data organisation Kaggele.com. The data contain the details of each crime case from 2010 to 2023. The following is a sample of the dataset.

DR_NO	Date Rptd	DATE OCC	TIME OC	AREA	AREA N	Rpt Dist N	Part 1-2	Crm Cd	Crm Cd D	Mocodes	Vict Age	Vict Sex	Vict Desc
10304468	1/8/2020 0:00	1/8/2020 0:00	2230	3	Southwest	377	2	624	BATTER	0444 0913	36	F	B
190101086	1/2/2020 0:00	1/1/2020 0:00	330	1	Central	163	2	624	BATTER	0416 1822	25	M	H
200110444	04/14/2020 12:00:00 AM	02/13/2020 12:0	1200	1	Central	155	2	845	SEX OFF	1501	0	X	X
191501505	1/1/2020 0:00	1/1/2020 0:00	1730	15	N Hollyw	1543	2	745	VANDAL	0329 1402	76	F	W
191921269	1/1/2020 0:00	1/1/2020 0:00	415	19	Mission	1998	2	740	VANDAL	329	31	X	X
200100501	1/2/2020 0:00	1/1/2020 0:00	30	1	Central	163	1	121	RAPE, FO	0413 1822	25	F	H
200100502	1/2/2020 0:00	1/2/2020 0:00	1315	1	Central	161	1	442	SHOPLIF	1402 2004	23	M	H
200100504	1/4/2020 0:00	1/4/2020 0:00	40	1	Central	155	2	946	OTHER M	1402 0392	0	X	X
200100507	1/4/2020 0:00	1/4/2020 0:00	200	1	Central	101	1	341	THEFT-G	1822 0344	23	M	B
201710201	06/19/2020 12:00:00 AM	05/26/2020 12:0	1925	17	Devonshir	1708	1	341	THEFT-G	1300 0202	0	X	X
200100509	1/4/2020 0:00	1/4/2020 0:00	2200	1	Central	192	1	330	BURGLA	1822 1414	29	M	A
200100510	1/5/2020 0:00	1/5/2020 0:00	955	1	Central	111	2	930	CRIMINA	0421 0906	35	M	O
200100514	1/5/2020 0:00	1/5/2020 0:00	1355	1	Central	162	1	341	THEFT-G	1822 0344	41	M	A
200100515	1/7/2020 0:00	1/7/2020 0:00	1638	1	Central	162	1	648	ARSON	1402 1501	0	X	X
200100520	1/8/2020 0:00	1/8/2020 0:00	1805	1	Central	128	1	442	SHOPLIF	0325 1402	24	F	H
211916029	11/26/2021 12:00:00 AM	11/30/2020 12:0	730	19	Mission	1916	2	626	INTIMAT	2000 1814	24	F	H
201116159	11/29/2020 12:00:00 AM	11/28/2020 12:0	2018	11	Northeast	1124	2	626	INTIMAT	0400 0416	34	F	H
200506268	02/22/2020 12:00:00 AM	02/22/2020 12:0	1900	5	Harbor	511	1	440	THEFT PI	0319 0344	29	F	W
210916801	11/22/2021 12:00:00 AM	11/19/2020 12:0	1200	9	Van Nuys	932	2	354	THEFT O	1501 1822	46	M	B
200100535	01/14/2020 12:00:00 AM	01/14/2020 12:0	1330	1	Central	152	1	210	ROBBER	0416 0411	66	M	B
200914517	9/10/2020 0:00	9/9/2020 0:00	1735	9	Van Nuys	909	2	354	THEFT O	0377 1822	40	M	O

Data Description

The following is the Data Description from the Open Data website, and this information may help you summarise the data dictionary for this project.

Data item	Explanation (where appropriate)
DR_NO	Official file number
DateRptd	Date of the Crime Reported
DATE OCC	The date of the Crime occurred
TIME OCC	Time of The crime occurred in 24 24-hour military time format
AREA NAME	The 21 Geographic Areas or Patrol Divisions are also given a name designation that references a landmark or the surrounding community that it is responsible
Crm Cd Desc	Description of the Crime
Vict Age	Age of the Victim
Weapon Desc	Weapon Used
LAT	Latitude of the Crime
LON	Longitude of the Crime

Task

Data Preparation

You will receive 13 Years of Crime data for your project, you have to ETL (using **R** or **Python**) before pouring the data into Tableau:

Stage 1

1. The dataset originally divided into two timelines, 2010 to 2019 and 2020 to now(2023). You must combine them to single dataset file.
2. The Date Data are messy, For example the 12:00:00 AM in the Report Date and occurred Date was useless, you have to extract the date from the date string.

Before

Date Rptd
1/8/2020 0:00
1/2/2020 0:00
04/14/2020 12:00:00 AM
1/1/2020 0:00
1/1/2020 0:00

After

Date.Rptd	DATE.OCC
20/2/2010	20/2/2010
13/9/2010	12/9/2010
9/8/2010	9/8/2010

3. Convert the occurred time military hour to standard time format and combine with occurred date

4. Before

DATE OCC	TIME OC
1/8/2020 0:00	2230
1/1/2020 0:00	330
02/13/2020 12:0	1200

5.

6. After

DATE.OCC	TIME.OC
20/2/2010	13:50
12/9/2010	0:45
9/8/2010	15:15+

8.

Occ_DateTime
20/2/2010 13:50
12/9/2010 0:45
9/8/2010 15:15

4. Fill “No weapon Recorded” to all missing value in Weapon Descrption

Before

Weapon Desc
STRONG-ARM (HANDS, FI
UNKNOWN WEAPON/OTH
LLING (APARTMENT, DUP
STORE
UNKNOWN WEAPON/OTH
ORE

After

Weapon.Desc
No Weapon Recorde
No Weapon Recorde
No Weapon Recorde

Stage 2 (Open end)

1. Select all necessary columns for your project

Before

ME OC	AREA	AREA Na	Rpt Dist N	Part 1-2	Crm Cd	Crm Cd D	Mocodes	Vict Age	Vict Sex	Vict Desc	Premis Cd	Premis De	Weapon U	Weapon Desc	Status	Status Des	Crm Cd 1	Crm Cd 2	Crm Cd 3	Crm Cd 4	LOCA
2230	3	Southwest	377	2	624	BATTER	0444 0913	36	F	B	501	SINGLE I	400	STRONG-ARM (HANDS, FI	AO	Adult Oth	624				1100
330	1	Central	163	2	624	BATTER	0416 1822	25	M	H	102	SIDEWAI	500	UNKNOWN WEAPON/OTH	IC	Invest Cor	624				700 S
1200	1	Central	155	2	845	SEX OFF	1501	0	X	X	726	POLICE FACILITY			AA	Adult Arre	845				200 E
1730	15	N Hollyw	1543	2	745	VANDALI	0329 1402	76	F	W	502	MULTI-UNIT DWELLING (APARTMENT, DUP			IC	Invest Cor	745	998			5400
415	19	Mission	1998	2	740	VANDAL	329	31	X	X	409	BEAUTY SUPPLY STORE			IC	Invest Cor	740				14400
30	1	Central	163	1	121	RAPE, FC	0413 1822	25	F	H	735	NIGHT C	500	UNKNOWN WEAPON/OTH	IC	Invest Cor	121	998			700 S
1315	1	Central	161	1	442	SHOPLIF	1402 2004	23	M	H	404	DEPARTMENT STORE			IC	Invest Cor	442	998			700 S
40	1	Central	155	2	946	OTHER M	1402 0392	0	X	X	726	POLICE FACILITY			IC	Invest Cor	946	998			200 E
200	1	Central	101	1	341	THEFT-G	1822 0344	23	M	B	502	MULTI-UNIT DWELLING (APARTMENT, DUP			IC	Invest Cor	341	998			700
1925	17	Devonshir	1708	1	341	THEFT-G	1300 0202	0	X	X	203	OTHER BUSINESS			AO	Adult Oth	341				11900
2200	1	Central	192	1	330	BURGLA	1822 1414	29	M	A	101	STREET	306	ROCK/THROWN OBJECT	IC	Invest Cor	330				15TH
955	1	Central	111	2	930	CRIMINA	0421 0906	35	M	O	108	PARKIN(	511	VERBAL THREAT	IC	Invest Cor	930				800 N
1355	1	Central	162	1	341	THEFT-G	1822 0344	41	M	A	503	HOTEL			AA	Adult Arre	341				800 S
1638	1	Central	162	1	648	ARSON	1402 1501	0	X	X	404	DEPARTI	500	UNKNOWN WEAPON/OTH	IC	Invest Cor	648	998			700 W
1805	1	Central	128	1	442	SHOPLIF	0325 1402	24	F	H	252	COFFEE SHOP (STARBUCKS, COFFEE BEAN,			IC	Invest Cor	442				100 S

After

DR_NO	Date.Rptd	DATE.OCC	TIME.OC	AREA.Na	Crm.Cd.D	Vict.Age	Vict.Sex	Vict.Desc	Premis.De	Weapon.Desc	Status.Des	LAT	LO
1307355	20/2/2010	20/2/2010	13:50	Newton	VIOLATI	48	M	H	SINGLE I	No Weapon Recorde	Adult Arre	33.9825	-
11401303	13/9/2010	12/9/2010	0:45	Pacific	VANDAL	0	M	W	STREET	No Weapon Recorde	Invest Cor	33.9599	-1
70309629	9/8/2010	9/8/2010	15:15	Newton	OTHER M	0	M	H	ALLEY	No Weapon Recorde	Invest Cor	34.0224	-1
90631215	5/1/2010	5/1/2010	1:50	Hollywoo	VIOLATI	47	F	W	STREET	HAND GUN	Invest Cor	34.1016	-
1E+08	3/1/2010	2/1/2010	21:00	Central	RAPE, A1	47	F	H	ALLEY	STRONG-ARM (HA	Invest Cor	34.0387	-1
1E+08	5/1/2010	4/1/2010	16:50	Central	SHOPLIF	23	M	B	DEPARTI	No Weapon Recorde	Adult Arre	34.048	-1
1E+08	8/1/2010	7/1/2010	20:05	Central	BURGLA	46	M	H	STREET	No Weapon Recorde	Invest Cor	34.0389	-1
1E+08	9/1/2010	8/1/2010	21:00	Central	ASSAUL	51	M	B	OTHER P	UNKNOWN WEAP	Adult Arre	34.0435	-1
1E+08	9/1/2010	9/1/2010	2:30	Central	ASSAUL	30	M	H	PARKIN(	STRONG-ARM (HA	Invest Cor	34.045	-1
1E+08	9/1/2010	6/1/2010	21:00	Central	THEFT-G	55	M	W	OTHER P	No Weapon Recorde	Invest Cor	34.0538	-1
1E+08	14/1/2010	14/1/2010	14:45	Central	BATTER	38	F	B	STREET	STRONG-ARM (HA	Invest Cor	34.064	-1

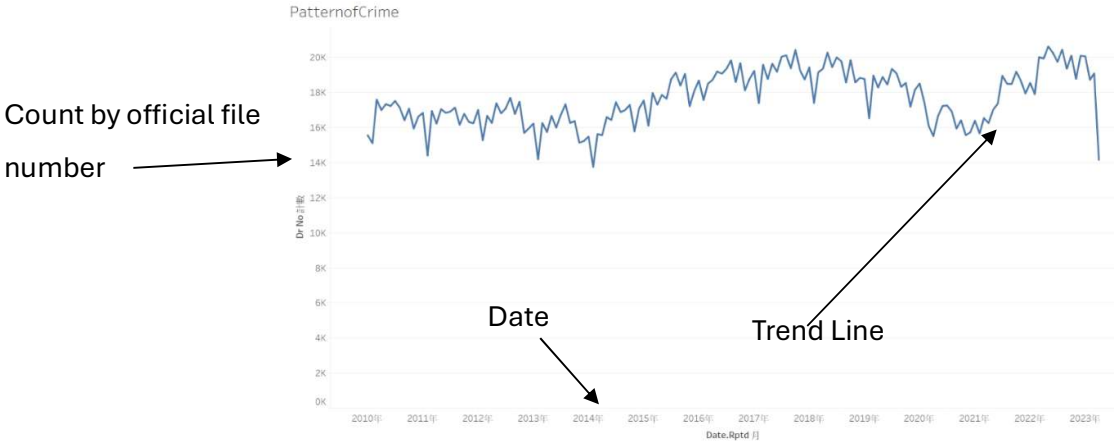
Stage 3

1. Save your cleared data in CSV or Excel format named "GroupName_Crime_Data".

Mandatory task

Visualize the following requirements, the below plotting is only for your reference.

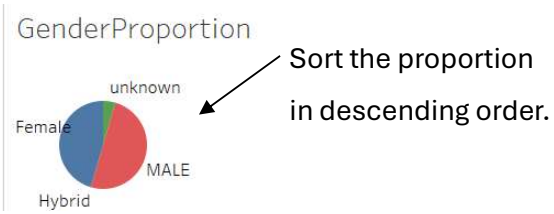
1. The time series trend line on the number of Crimes among years



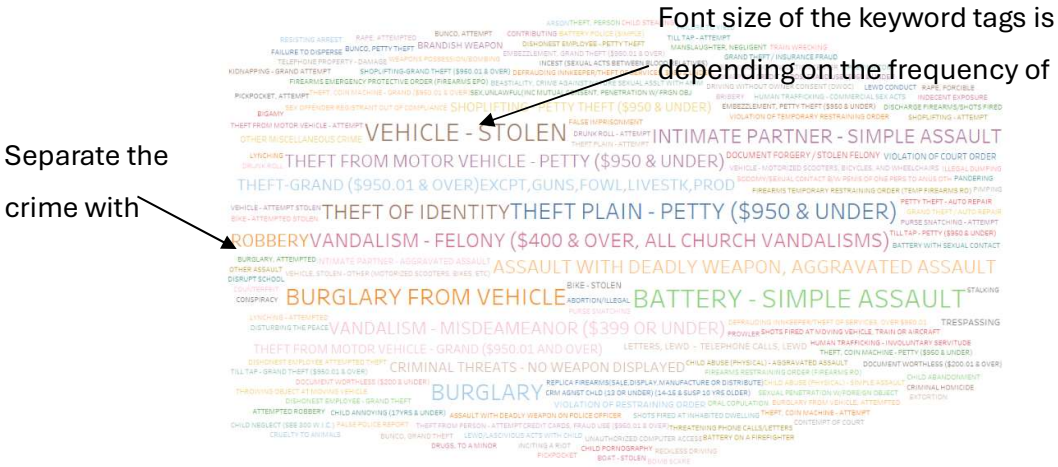
2. Proportion of Victim gender.

Follow the below table to convert the gender code to the description by Tableau calculated field

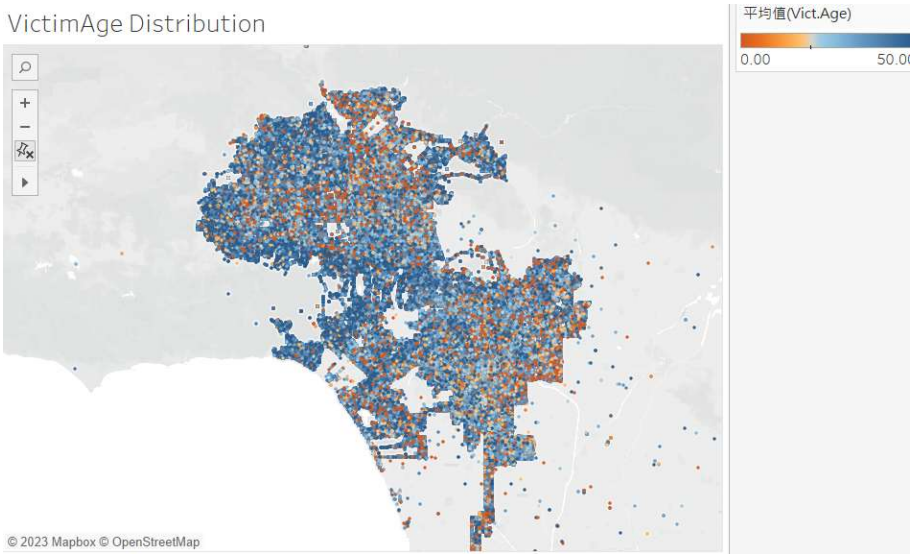
Code	Description
F	Female
H	Hybrid
M	Male
N	No tell
X,-,Null	unknown



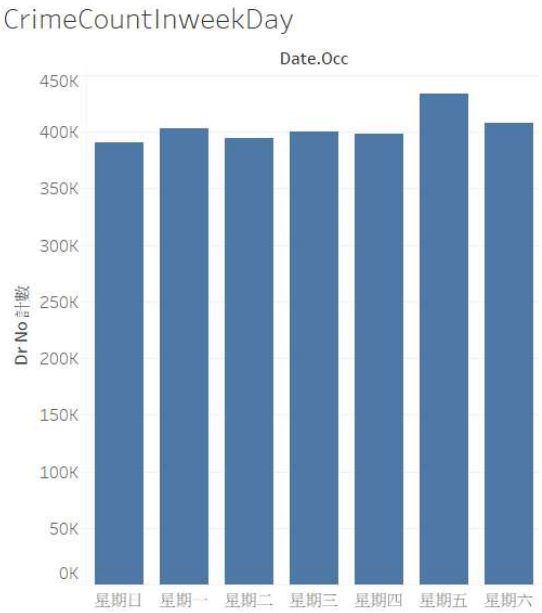
3. Crime Keyword Cloud



4. Crime Map (coloured by Age)

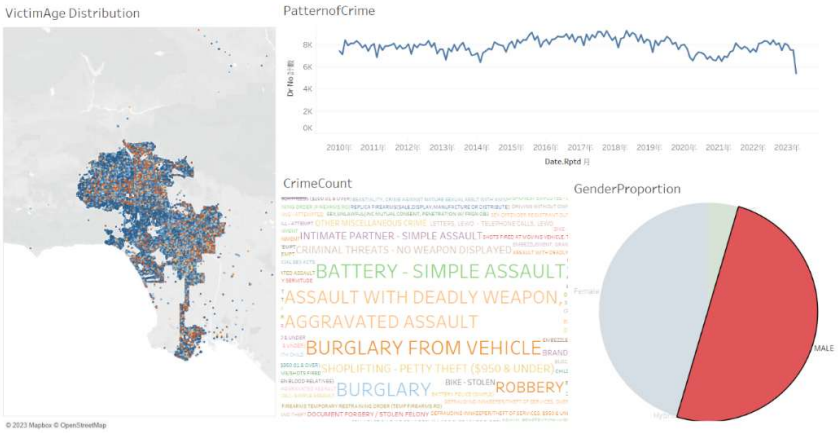


5. Crime count by weekday (additional mark for visualization average count by using tableau only)



Advanced technique task

This task is a freestyle from your idea. First, you have to create dashboards for your project purpose. Remember that you must complete one of the dashboards by using visualised plottings from the last section and drilling some valuable findings. Furthermore, you still have to plot some graphs not mentioned in the previous section and consolidate them as a story-telling dashboard. Below is a dashboard for your reference. You may follow the guideline of Steve Few's informative dashboards for designing a good dashboard. Remember, a good dashboard is not an ornament; it should be a tool for assisting your findings. PLEASE DON'T USE A SINGLE TYPE OF GRAPH FOR YOUR DASHBOARD.



Finding demonstrations

You should present your idea/findings through the story mode in 15 mins. No PowerPoint or other presentation software is allowed. The presentation should make use of the following Tableau features:

1. Story mode
2. Dashboard
3. Filtering data
4. Drill-in data
5. Story-telling
6. Presentation Mode
7. Single Graph sheet

Furthermore, you also have prepared a report for our record. You may find the report template in Moodle. The Report have to include the following parts:

1. Introduction of project
2. Data Dictionary
3. Findings
4. Suggestions
5. Reference (if any)

Project Case 2425 Full Document

You are required to form a group of 2~ 3 classmates or even work individually. Each group or individual will receive one set of data. All groups or individuals have to process the dataset and generate some findings. The data is about the transaction data of British Property. All groups and individuals use the provided data and finish the following tasks: Data preparation, Mandatory task, Advanced technique task, and Finding demonstrations. All tasks and demonstrations are required to utilise the features of Tableau Public and R for data pre-processing.

To Conduct the project smoothly, The project topic should not be limited to the following suggestions but should included in the theme of "**When Britain parallel with Pandemic**"**:**

5. Which type of property is popular in the UK during the Pandemic?
6. How is the pandemic affecting property transactions?
7. Which kind of property will become the next trend?

Dataset Description

Transaction record from the Government

This is public data from the UK Government Open Data website[4]. The data contain the details of each transaction, from 1995 to 2024. The following is a sample of the dataset.

T id	Price	Dateof Trans	Postalcode	Type	Old New	Duration	Paon	Saon	Street	Locality	Town City	District
{A2479555-4F39-74C...	250,000	7/2/2020 0:00:00	EN8 8SA	Maisonettes	Established residential	Leasehold	EDWARDS COURT	FLAT 18	TURNERS HILL	null	WALTHAM CROSS	BROXBOURNE
{A2479555-4F3A-74C...	370,000	14/2/2020 0:00:00	WD6 1UG	Terraced	Established residential	Freehold	4	null	WELBECK CLOSE	null	BOREHAMWOOD	HERTSMERE
{A2479555-4F3B-74C...	685,000	6/3/2020 0:00:00	AL6 0PW	Detached	Established residential	Freehold	13	null	THE AVENUE	null	WELWYN	WELWYN HATFIELD
{A2479555-4F3C-74C...	133,000	10/1/2020 0:00:00	SG1 4EL	Maisonettes	Established residential	Leasehold	384	null	YORK ROAD	null	STEVENAGE	STEVENAGE
{A2479555-4F3D-74C...	223,000	28/2/2020 0:00:00	SG2 7QR	Terraced	Established residential	Freehold	14	null	CONIFER WALK	null	STEVENAGE	STEVENAGE
{A2479555-4F3E-74C...	725,000	21/2/2020 0:00:00	SG13 7TQ	Detached	Established residential	Freehold	2	null	HONEYSUCKLE CLOSE	null	HERTFORD	EAST HERTFORDSHIF
{A2479555-4F3F-74C...	2,075,000	28/2/2020 0:00:00	AL1 4BG	Detached	Established residential	Freehold	23	null	HOMEWOOD ROAD	null	ST ALBANS	ST ALBANS
{A2479555-4F40-74C...	708,200	25/2/2020 0:00:00	CM23 5TB	Detached	Established residential	Freehold	24	null	WRAGLINGS	null	BISHOP'S STORTFORD	EAST HERTFORDSHIF
{A2479555-4F41-74C...	258,000	28/2/2020 0:00:00	WD6 3HL	Maisonettes	Established residential	Leasehold	VALENTINE COURT	3	BARNET LANE	ELSTREE	BOREHAMWOOD	HERTSMERE
{A2479555-4F42-74C...	270,000	20/2/2020 0:00:00	HP1 2TW	Terraced	Established residential	Freehold	78	null	THE PASTURES	null	HEMEL HEMPSTEAD	DACORUM
{A2479555-4F43-74C...	557,500	21/2/2020 0:00:00	SG5 1UB	Terraced	Established residential	Freehold	139	null	BEARTON ROAD	null	HITCHIN	NORTH HERTFORDS.
{A2479555-4F44-74C...	295,000	20/2/2020 0:00:00	HP2 6LL	Terraced	Established residential	Freehold	28	null	CLAYMORE	null	HEMEL HEMPSTEAD	DACORUM
{A2479555-4F45-74C...	319,995	28/2/2020 0:00:00	EN8 8TS	Terraced	Established residential	Freehold	11	null	RUSSELLS RIDE	CHESHUNT	WALTHAM CROSS	BROXBOURNE
{A2479555-4F46-74C...	265,500	10/2/2020 0:00:00	CM23 2AJ	Terraced	Established residential	Freehold	7	null	CHERRY GARDENS	null	BISHOP'S STORTFORD	EAST HERTFORDSHIF

[4] Data source website: https://www.gov.uk/government/statistical-data-sets/price-paid-data-downloads#single-file

Data Description

The following is the Data Description from the UK Open data website, and this information may help you summarise the data dictionary for this project.

Data item	Explanation (where appropriate)
Transaction unique identifier	A reference number which is generated automatically recording each published sale. The number is unique and will change each time a sale is recorded.
Price	Sale price stated on the transfer deed.
Date of Transfer	Date when the sale was completed, as stated on the transfer deed.
Postcode	This is the postcode used at the time of the original transaction. Note that postcodes can be reallocated and these changes are not reflected in the Price Paid Dataset.
Property Type	D = Detached, S = Semi-Detached, T = Terraced, F = Flats/Maisonettes, O = Other Note that: - we only record the above categories to describe property type, we do not separately identify bungalows - end-of-terrace properties are included in the Terraced category above - 'Other' is only valid where the transaction relates to a property type that is not covered by existing values, for example where a property comprises more than one large parcel of land
Old/New	Indicates the age of the property and applies to all price paid transactions, residential and non-residential. Y = a newly built property, N = an established residential building

Data item	Explanation (where appropriate)
Duration	Relates to the tenure: F = Freehold, L= Leasehold etc. Note that HM Land Registry does not record leases of 7 years or less in the Price Paid Dataset.
PAON	Primary Addressable Object Name. Typically the house number or name.
SAON	Secondary Addressable Object Name. Where a property has been divided into separate units (for example, flats), the PAON (above) will identify the building and a SAON will be specified that identifies the separate unit/flat.
Street	
Locality	
Town/City	
District	
County	
PPD Category Type	Indicates the type of Price Paid transaction. A = Standard Price Paid entry, includes single residential property sold for value. B = Additional Price Paid entry including transfers under a power of sale/repossessions, buy-to-lets (where they can be identified by a Mortgage), transfers to non-private individuals and sales where the property type is classed as 'Other'. Note that category B does not separately identify the transaction types

Data item	Explanation (where appropriate)
	stated. HM Land Registry has been collecting information on Category A transactions from January 1995. Category B transactions were identified from October 2013.
Record Status - monthly file only	Indicates additions, changes and deletions to the records.(see guide below). A = Addition C = Change D = Delete Note that where a transaction changes category type due to misallocation (as above) it will be deleted from the original category type and added to the correct category with a new transaction unique identifier.

Task

Data preparation

You will receive 29,145,919 transaction records for your project, and you have to ETL before pouring the data into the Tableau workbook. Therefore, create an R notebook to:

8. Extract the data from the pandemic period to now (The year 2020 to the Year 2024).
9. Since the dataset does not contain headers, you must label all headers based on the information provided.
10. You have to convert all Category codes to Category labels.
11. Save your result to a CSV format file named Housing_Final.csv
12. Add the Label “After pandemic” to the record that the transaction was conducted after Year 2022, otherwise, the label is “During pandemic”

Deliverable:

2. Several dataset files in a CSV format:
    I. Main Data should be in one file (Housing_Final.csv)

II. Supplementary data should separate data files (HousingLocation.csv)

Hint:

6. Filtrate all unnecessary columns (up to you)
7. Postal Code can help you to find all property geographical data.
8. The below is just for your reference (in a CSV format)

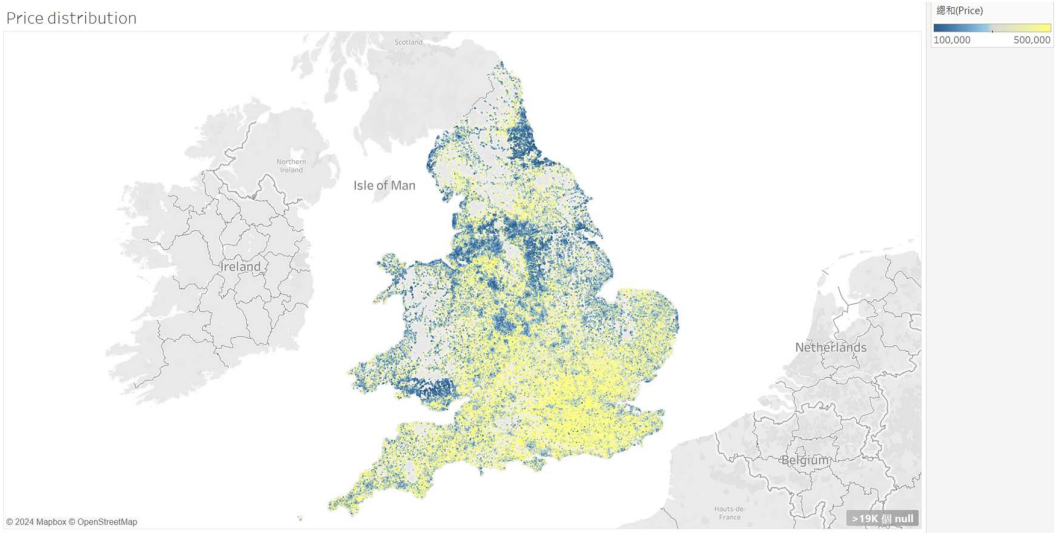
T id	Price	Dateof Trans	Postalcode	Type	Old New	Duration	Paon	Saon	Street	Locality	Town City	District
{A2479555-4F39-74C...	250,000	7/2/2020 0:00:00	EN8 8SA	Maisonettes	Established residential	Leasehold	EDWARDS COURT	FLAT 18	TURNERS HILL		WALTHAM CROSS	BROXBOURNE
{A2479555-4F3A-74C...	370,000	14/2/2020 0:00:00	WD6 1UG	Terraced	Established residential	Freehold	4		WELBECK CLOSE		BOREHAMWOOD	HERTSMERE
{A2479555-4F3B-74C...	685,000	6/3/2020 0:00:00	AL6 0PW	Detached	Established residential	Freehold	13		THE AVENUE		WELWYN	WELWYN HATFIELD
{A2479555-4F3C-74C...	133,000	10/1/2020 0:00:00	SG1 4EL	Maisonettes	Established residential	Leasehold	384		YORK ROAD		STEVENAGE	STEVENAGE
{A2479555-4F3D-74C...	223,000	28/2/2020 0:00:00	SG2 7QR	Terraced	Established residential	Freehold	14		CONIFER WALK		STEVENAGE	STEVENAGE
{A2479555-4F3E-74C...	725,000	21/2/2020 0:00:00	SG13 7TQ	Detached	Established residential	Freehold	2		HONEYSUCKLE CLOSE		HERTFORD	EAST HERTFORDSHIF
{A2479555-4F3F-74C...	2,075,000	28/2/2020 0:00:00	AL1 4BG	Detached	Established residential	Freehold	23		HOMEWOOD ROAD		ST ALBANS	ST ALBANS
{A2479555-4F40-74C...	708,200	25/2/2020 0:00:00	CM23 5TB	Detached	Established residential	Freehold	24		WRAGLINGS		BISHOP'S STORTFORD	EAST HERTFORDSHIF
{A2479555-4F41-74C...	258,000	28/2/2020 0:00:00	WD6 3HL	Maisonettes	Established residential	Leasehold	VALENTINE COURT	3	BARNET LANE	ELSTREE	BOREHAMWOOD	HERTSMERE
{A2479555-4F42-74C...	270,000	20/2/2020 0:00:00	HP1 2TW	Terraced	Established residential	Freehold	78		THE PASTURES		HEMEL HEMPSTEAD	DACORUM
{A2479555-4F43-74C...	557,500	21/2/2020 0:00:00	SG5 1UB	Terraced	Established residential	Freehold	139		BEARTON ROAD		HITCHIN	NORTH HERTFORDS.
{A2479555-4F44-74C...	295,000	20/2/2020 0:00:00	HP2 6LL	Terraced	Established residential	Freehold	28		CLAYMORE		HEMEL HEMPSTEAD	DACORUM
{A2479555-4F45-74C...	319,995	28/2/2020 0:00:00	EN8 8TS	Terraced	Established residential	Freehold	11		RUSSELLS RIDE	CHESHUNT	WALTHAM CROSS	BROXBOURNE
{A2479555-4F46-74C...	265,500	10/2/2020 0:00:00	CM23 2AJ	Terrac					CHERRY GARDENS		BISHOP'S STORTFORD	EAST HERTFORDSHIF

Postalcode1	Lat	Long
EN8 8SA	51.70032	-0.03356
WD6 1UG	51.65829	-0.27874
AL6 0PW	51.84216	-0.20448
SG1 4EL	51.92137	-0.18578
SG2 7QR	51.91065	-0.16237
SG13 7TQ	51.79779	-0.05456
AL1 4BG	51.76257	-0.31302
CM23 5TB	51.86246	0.17150

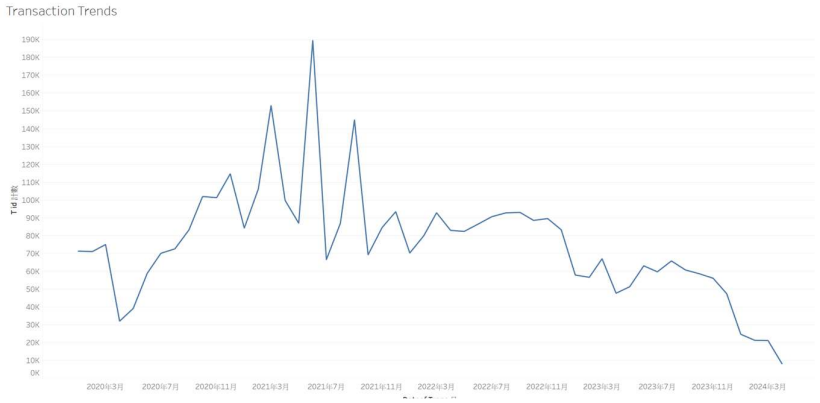
Mandatory task

Visualise the following requirements

8. The property price distribution in the UK.

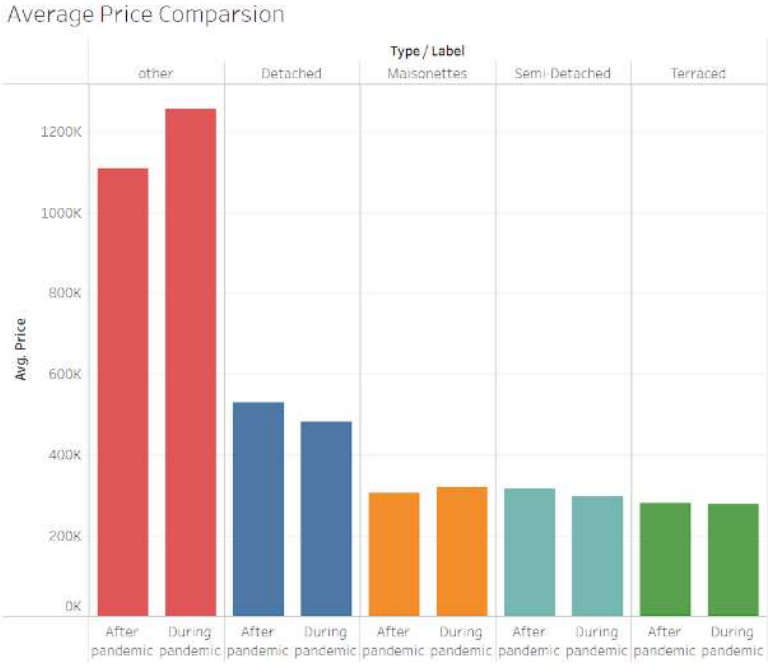


9. The trend of the number of transactions.

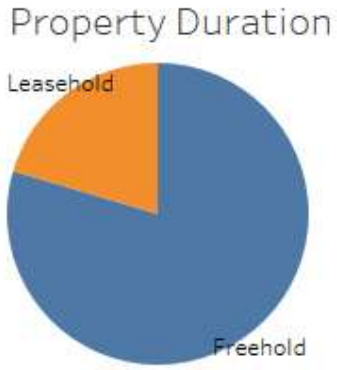


10. A City Keyword cloud measured by media house price coloured by levels of average house price (Show the highest 100 cities)

11. Average Price Comparison of each Type of Property based on pandemic period

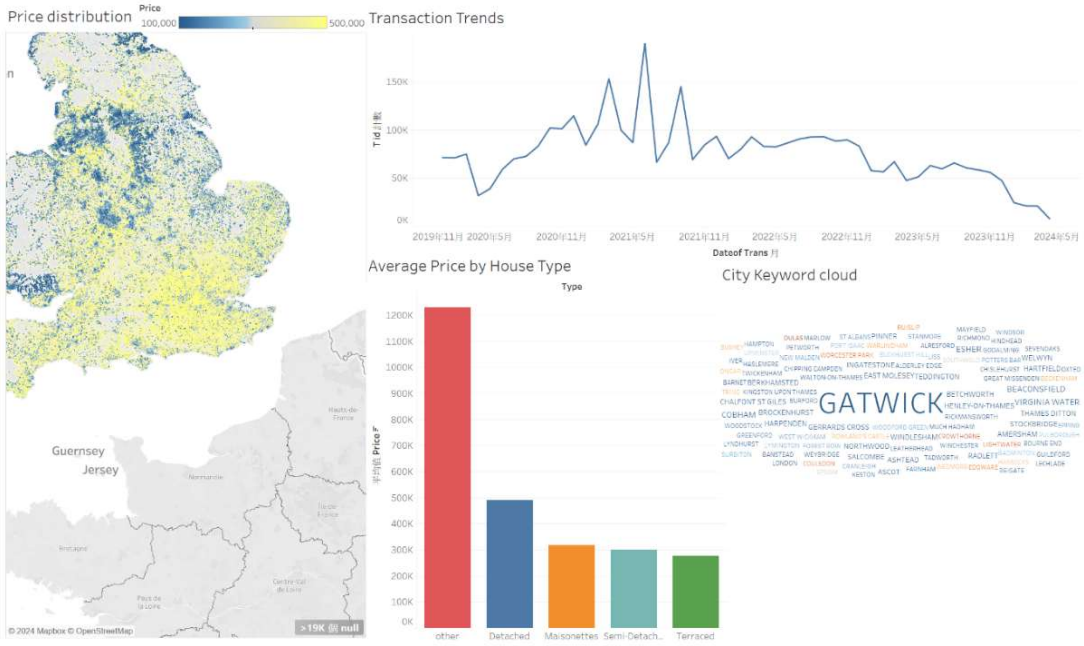


12. The proportion of Property Duration types.



Advanced technique task

This task is a freestyle from your idea. First, you have to create dashboards for your project purpose. Bear in mind that you must complete one of the dashboards by using visualized plottings from the last section and drilling some valuable findings. Furthermore, you still have to plot some graphs not mentioned in the previous section and consolidate them as a story-telling dashboard. The below is a dashboard for your reference. You may follow the guidelines of Steve Few informative dashboards for designing a good dashboard. Remember, a good dashboard is not an ornament; it should be a tool for assisting your findings. PLEASE DON'T USE A SINGLE TYPE OF GRAPH FOR YOUR DASHBOARD.



Finding demonstrations

You should present your idea/findings through the story mode in 10 minutes. No PowerPoint or other presentation software is allowed. The presentation should make use of the following Tableau features:

1. Story mode
2. Dashboard
3. Filtering data
4. Drill-in data
5. Story-telling
6. Presentation Mode
7. Single Graph sheet

Furthermore, you also have prepared a report for our record. You may find the report template in Moodle. The Report has to include the following parts:

1. Introduction of project
2. Data Dictionary
3. Findings
4. Suggestions
5. Reference (if any)

Project Case 2526 Full Document

The data is about the Top 50 Popular Songs in 73 Countries. All groups and individuals use the provided data and finish the following tasks: Data preparation, Mandatory task, Advanced technique task, and Finding demonstrations. All tasks and demonstrations are required to utilise the features of Tableau Public and R for data pre-processing.

To Conduct the project smoothly, The project topic should not be limited to the following suggestions but should included in the theme of "**Music with You**"**:**

5. Review of song style in different country?
6. Relatship between artist and song?
7. Which kind of song will become the next trend?
8. The style of different artists.

Main Dataset Description

Popular Music Data

This is public data from the popular Music Platform[5]. The data contain the details of each transaction, from 1995 to 2024. The following is a sample of the dataset.

2RkZ5LkI	Ordinary	Alex Warren	1	1	0	2025/6/11	95	FALSE	186964	You'll Be	2024/9/26	0.368	0.694	2	-6.141	1	0.06
42UBPzR	Manchild	Sabrina Carpenter	2	-1	48	2025/6/11	89	TRUE	213645	Manchild	2025/6/5	0.731	0.685	7	-5.087	1	0.0572
0FTmksd	back to friends	sombr	3	0	1	2025/6/11	98	FALSE	199032	back to fri	2024/12/27	0.436	0.723	1	-2.291	1	0.0301
7so0lgd0z	Die With A Smile	Lady Gaga, Bruno	4	0	-1	2025/6/11	91	FALSE	251667	MAYHEM	2025/3/7	0.519	0.601	6	-7.727	0	0.0317
6dOtVTDd	BIRDS OF A FEAT	Billie Eilish	5	1	0	2025/6/11	100	FALSE	210373	HIT ME F	2024/5/17	0.747	0.507	2	-10.171	1	0.0358
27xkOIEF	Don? Say You Lo	Jin	6	-1	-4	2025/6/11	93	FALSE	180716	Echo	2025/5/16	0.729	0.562	8	-5.49	1	0.0304
6iOndD4C	La Plena - W Sound	W Sound, Be酸le,	7	0	0	2025/6/11	95	TRUE	150001	La Plena (	2025/2/19	0.894	0.643	5	-3.485	1	0.132
4AajxCEv	undressed	sombr	8	0	-2	2025/6/11	96	FALSE	182088	undressed	2025/3/21	0.642	0.887	0	-3.883	1	0.0406
4wJ5Qq0j	APT.	ROS?, Bruno Mars	9	1	-1	2025/6/11	89	FALSE	169917	rosie	2024/12/6	0.777	0.783	0	-4.477	0	0.26
4WFgvK\	Love Me Not	Ravyn Lenae	10	-1	9	2025/6/11	91	FALSE	213466	Bird's Eye	2024/8/9	0.779	0.731	0	-6.176	1	0.0306
3QaPy1Kg	WILDFLOWER	Billie Eilish	11	1	1	2025/6/11	96	FALSE	261466	HIT ME F	2024/5/17	0.467	0.247	6	-12.002	0	0.0431
6eLQXa6(	TU SANCHO	Fuerza Regida	12	2	-1	2025/6/11	90	TRUE	177994	111XPAN	2025/5/2	0.689	0.84	8	-6.423	1	0.0457
2yWlGEg	Just Keep Watching	Tate McRae, F1 Th	13	0	-3	2025/6/11	91	FALSE	142550	Just Keep	2025/5/30	0.656	0.796	9	-5.856	0	0.113
3xkHsmp(	Beautiful Things	Benson Boone	14	3	-1	2025/6/11	93	FALSE	180304	Fireworks	2024/4/5	0.472	0.471	10	-5.692	1	0.0603
2HRqTpk	Espresso	Sabrina Carpenter	15	1	6	2025/6/11	91	TRUE	175459	Short n' S	2024/8/23	0.699	0.776	0	-5.282	1	0.0293
21IYMdz	Sailor Song	Gigi Perez	16	-1	-2	2025/6/11	89	FALSE	211978	At The Be	2025/4/25	0.499	0.392	11	-10.441	1	0.0262
3sK8wGT	DtMF	Bad Bunny	17	2	-2	2025/6/11	97	TRUE	237117	DeB? TiR	2025/1/5	0.625	0.131	7	-27.405	0	0.0717
7tI8dRuH	Who	Jimin	18	-7	-9	2025/6/11	90	FALSE	170887	MUSE	2024/7/19	0.66	0.756	0	-3.743	0	0.032
7ne4VBA	That? So True	Gracie Abrams	19	2	-1	2025/6/11	96	TRUE	166300	The Secre	2024/10/18	0.554	0.808	1	-4.169	1	0.0368
0QCIpQV	Shake It To The Ma	MOLIY, Silent Ad	20	-2	-4	2025/6/11	93	TRUE	178775	Shake It T	2025/2/21	0.856	0.548	0	-6.609	0	0.123

[5] Data source website: https://www.kaggle.com/datasets/asaniczka/top-spotify-songs-in-73-countries-daily-updated

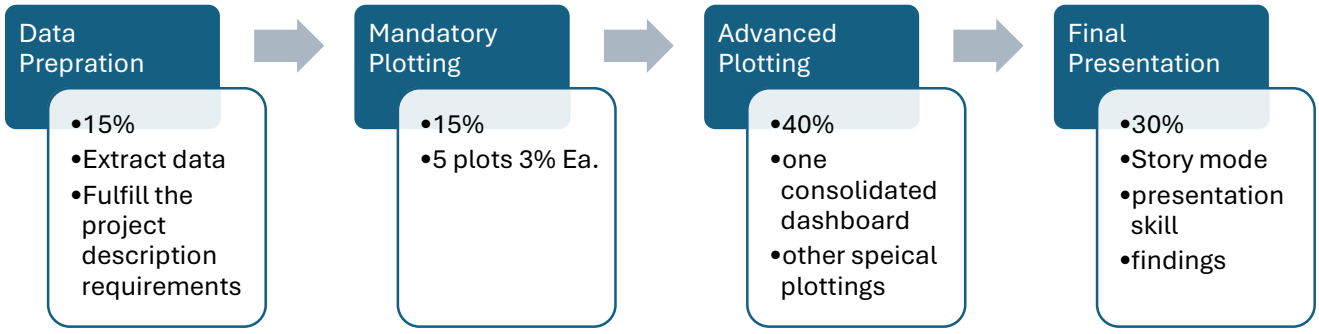
Data Description

The following is the initial Data Description of the dataset provided by Kaggle platform, and this information may help you summarise the finalized data dictionary for this project.

Column	Explanation (where appropriate)
spotify_id	The unique identifier for the song in the Spotify database
name	The title of the song
daily_rank	The daily rank of the song in the top 50 list
artists	The name(s) of the artist(s) associated with the song
daily_movement	The change in rankings compared to the previous day
weekly_movement	The change in rankings compared to the previous week
country	The ISO code of the country of the Top 50 Playlist. If Null, then the playlist if 'Global Top 50'
snapshot_date	The date on which the data was collected from the Spotify API
popularity	A measure of the song's current popularity on Spotify
is_explicit	Indicates whether the song contains explicit lyrics
duration_ms	The duration of the song in milliseconds

Column	Explanation (where appropriate)
album_name	The title of the album the song belongs to
album_release_date	The release date of the album the song belongs to
danceability	A measure of how suitable the song is for dancing based on various musical elements
energy	A measure of the intensity and activity level of the song.
key	The key of the song
loudness	The overall loudness of the song in decibels
mode	Indicates whether the song is in a major or minor key
speechiness	A measure of the presence of spoken words in the song
acousticness	A measure of the acoustic quality of the song
instrumentalness	A measure of the likelihood that the song does not contain vocals
liveness	A measure of the presence of a live audience in the recording
valence	A measure of the musical positiveness conveyed by the song
tempo	The tempo of the song in beats per minute
time_signature	The estimated overall time signature of the song

Task



Totally: 100%

Data preparation

You will receive 2,110,316 records for your project, and you have to ETL before pouring the data into the Tableau workbook. Therefore, create an R notebook to:

13. Extract the full name of ISO Country code from https://help.adjust.com/en/article/countries-by-region
14. convert the country to the fill name
15. Label the valence and speechiness value by the following table

Category/Range	< 1st quartile	1st quartile ~ Median	Median ~ 3rd quartile	> 3rd quartile
speechiness	Pure music	Song and words balaced	Word leads music	Words oriented
valence	darkness	Gray feeling	positive	brightness

16. Extract the Managed data to csv format and named Music_Final.csv

Deliverable:

3. Several dataset files in a CSV format:
   I. Main Data should be in one file (Music_Final.csv)

Hint:

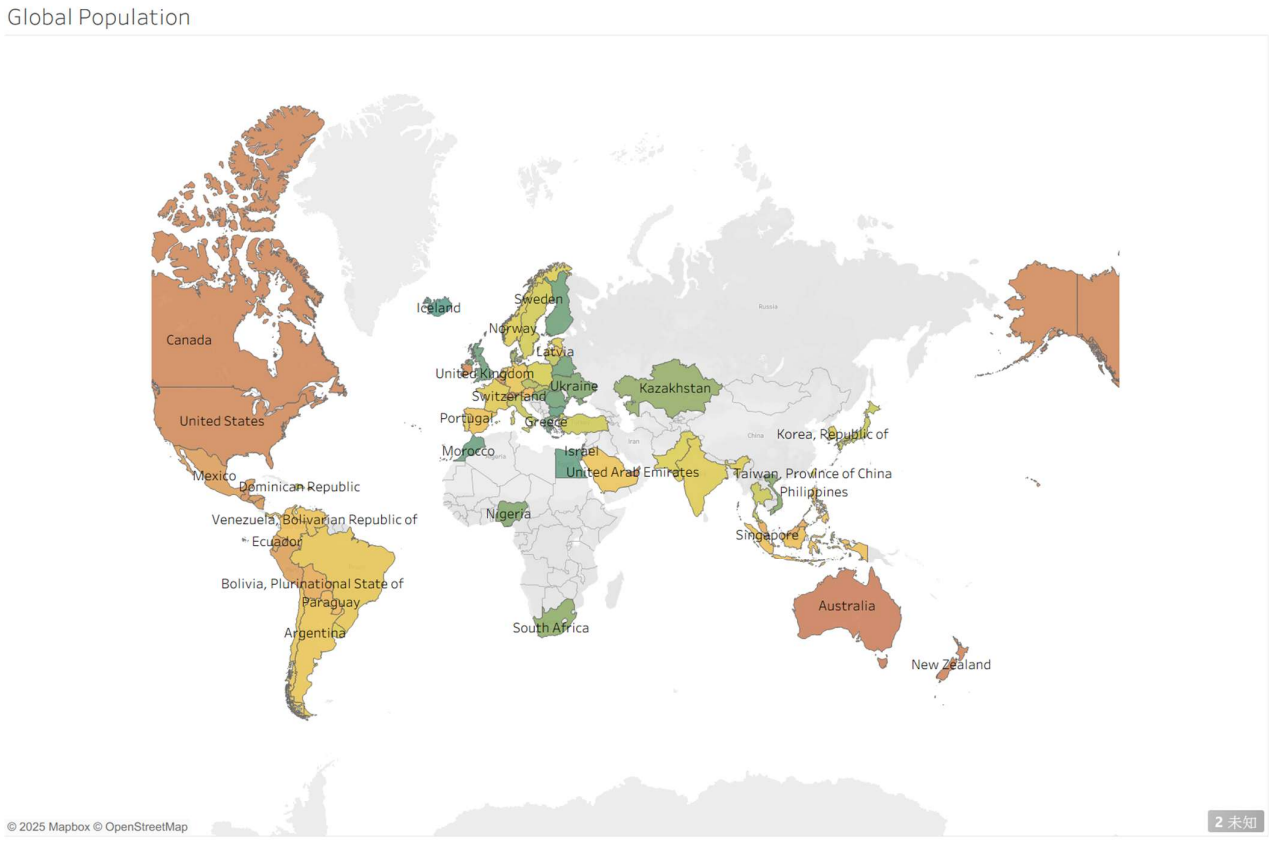
9. Filtrate all unnecessary columns (up to you)
10. The below is just for your reference (in a CSV format)

Name	country	spotify_id	name	artists	daily_rank	daily_mov	weekly_m	snapshot_c	popularity	is_explicit	duration_n	album_nar	album_relea	danceabilit	energy	key	loudness	mode	speechines	speechines	acousticne	instrument	liveness	valence	valence_la	tempo	time_signa	Region_name
Argentina	AR	5AkLEnor	Si Un D翻	Big One, I	41	1	-7	2025/3/1	69	FALSE	164021	Si Un D翻	2024/9/3	0.63	0.507	1	-4.865	1	0.0647	word leads	0.0386	0	0.0666	0.744	brightness	183.896	4	Latin America
Argentina	AR	6MuIUSY	Ese Maldit	No Te Va	36	5	0	2025/6/9	73	FALSE	185866	El Calor D	2012/10/18	0.643	0.616	11	-6.358	0	0.0266	pure music	0.469	5.74e-05	0.0886	0.652	positive	139.931	4	Latin America
Argentina	AR	38EGs4h8	Con otra	Cazzu	6	0	0	2025/4/17	85	FALSE	231016	Con otra	2025/3/19	0.581	0.61	10	-5.787	0	0.0492	song and v	0.274	0	0.949	0.55	positive	168.162	4	Latin America
Argentina	AR	7d6yK8v8	VeLD?	Bad Bunny	9	-2	-1	2025/4/17	93	TRUE	235336	DeB? TiR	2025/1/5	0.613	0.981	1	-18.638	1	0.149	words orie	0.29	0.0023	0.437	0.0337	darkness	101.852	3	Latin America
Argentina	AR	6WatFBL'	Si Antes T	KAROL G	12	-1	-6	2025/1/16	92	FALSE	195824	Si Antes T	2024/6/21	0.924	0.668	11	-6.795	1	0.0469	song and v	0.446	0.000594	0.0678	0.787	brightness	128.027	4	Latin America
Argentina	AR	4xXvGEi7	COLOCA	Renzo ED	23	4	-1	2025/3/31	73	FALSE	176250	COLOCA	2024/10/24	0.857	0.701	0	-3.483	0	0.063	word leads	0.434	0	0.0395	0.812	brightness	96.046	4	Latin America
Argentina	AR	5MKbWa	Amor de \	La T y La	7	1	0	2025/4/17	81	FALSE	188423	Amor de \	2024/8/23	0.65	0.507	2	-7.27	1	0.111	word leads	0.706	2.47e-06	0.0725	0.603	positive	179.984	4	Latin America
Argentina	AR	5DSMm0I	DOCTOR	Luck Ra, I	43	4	-1	2025/3/1	69	FALSE	144148	DOCTOR	2024/8/29	0.651	0.762	9	-2.969	1	0.0315	pure music	0.0834	0	0.0701	0.721	positive	90.989	4	Latin America
Argentina	AR	5QjmUqgI	UWAIE	Kapo	26	-1	-3	2024/10/1	89	FALSE	172427	UWAIE	2024/8/15	0.705	0.783	9	-4.783	0	0.0403	song and v	0.138	0	0.0984	0.454	gray Feelli	103.001	4	Latin America
Argentina	AR	4YnmQG	WYA REI	J Abdiel, I	30	-1	-2	2024/10/1	83	TRUE	321741	WYA REI	2024/5/10	0.733	0.527	0	-5.985	0	0.0486	song and v	0.364	0	0.0999	0.371	gray Feelli	110.045	4	Latin America
Argentina	AR	0qeIFEGR	DEGENEI	Myke Tow	43	-1	-6	2025/3/21	84	TRUE	130880	LA PANT	2024/10/17	0.744	0.709	6	-5.778	0	0.284	words orie	0.0501	0	0.0847	0.712	positive	194.055	4	Latin America
Argentina	AR	6Es8Sk3x	La_Origin	Emilia, TI	25	-1	0	2024/10/1	76	FALSE	140625	.mp3	2023/11/3	0.833	0.809	5	-2.751	0	0.0381	pure music	0.0836	0	0.112	0.905	brightness	128.055	4	Latin America
Argentina	AR	0SoBfYjQ	Hoy	Valentino	3	0	1	2024/6/25	71	FALSE	198495	Hoy	2024/5/22	0.451	0.711	11	-4.612	1	0.0426	song and v	0.24	0	0.154	0.661	positive	176.161	4	Latin America
Argentina	AR	2RmtEc22	Ojos Verd	Nicki Nicc	4	0	2	2024/6/25	77	FALSE	142410	Ojos Verd	2024/4/24	0.582	0.834	4	-2.654	1	0.0394	song and v	0.224	2.67e-06	0.145	0.95	brightness	167.945	4	Latin America
Argentina	AR	3sra6aEW	IMAN (Tv	Maria Bec	5	5	4	2024/6/25	75	FALSE	122405	IMAN (Tv	2024/5/16	0.8	0.679	3	-5.01	0	0.0485	song and v	0.154	2.59e-05	0.205	0.974	brightness	106.045	4	Latin America
Argentina	AR	4dLLdfiI	Perdonarte	Los ?ngele	2	0	0	2024/6/25	80	FALSE	193933	Perdonarte	2024/5/23	0.672	0.702	11	-4.662	0	0.0314	pure music	0.168	0	0.0445	0.615	positive	90.039	4	Latin America
Argentina	AR	5rQSQlZX	Un Besito	Salastkbro	7	-1	-2	2024/6/25	74	FALSE	160750	Un Besito	2024/3/14	0.693	0.641	2	-5.775	0	0.0359	pure music	0.368	0	0.294	0.514	gray Feelli	92.01	4	Latin America
Argentina	AR	5AGnkbW	Luck Ra: I	Bizarrap, I	6	0	-2	2025/1/30	81	FALSE	170953	Luck Ra: I	2024/12/27	0.704	0.89	0	-4.649	0	0.13	words orie	0.0623	0	0.0982	0.638	positive	159.927	4	Latin America
Argentina	AR	2xbH0CuI	SINVERG	Emanero,	27	0	-3	2024/10/1	72	FALSE	217490	SINVERG	2023/12/5	0.764	0.738	0	-4.774	1	0.0336	pure music	0.182	0	0.173	0.647	positive	93.963	4	Latin America
Argentina	AR	3cWo162\	Tu Foto - I	Q' Lokura,	6	-1	1	2024/6/25	72	FALSE	180385	Tu Foto	2024/5/16	0.561	0.803	11	-5.132	0	0.073	word leads	0.146	0	0.188	0.338	darkness	153.669	4	Latin America

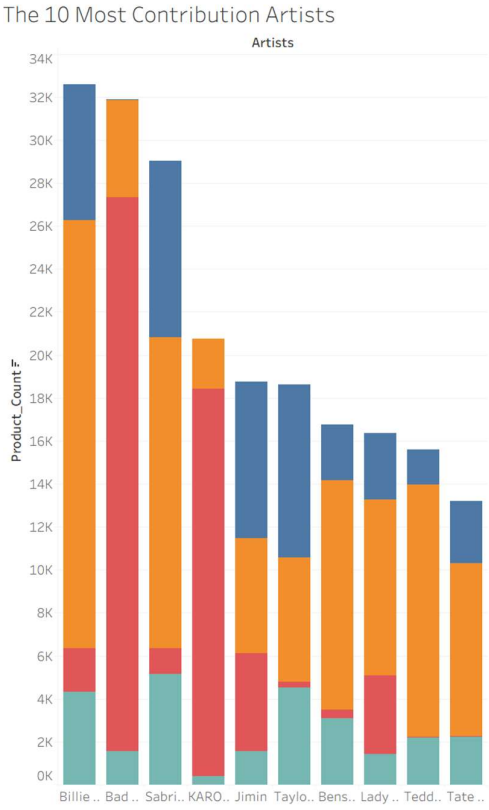
Mandatory task (2% Ea.)

Visualise the following requirements

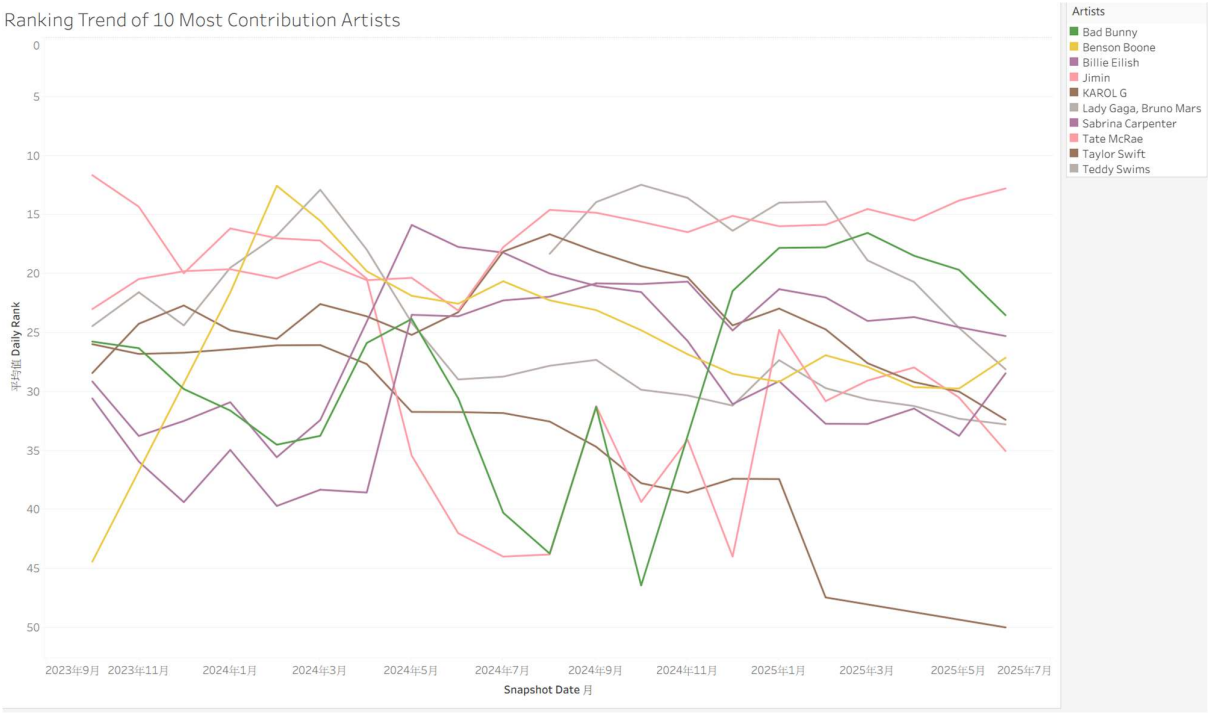
11. The global population coloured by value of population value.



12. The 10 most contribution artists and bar should display the component of regions.



13. The rank trend of 10 most contribution artists.

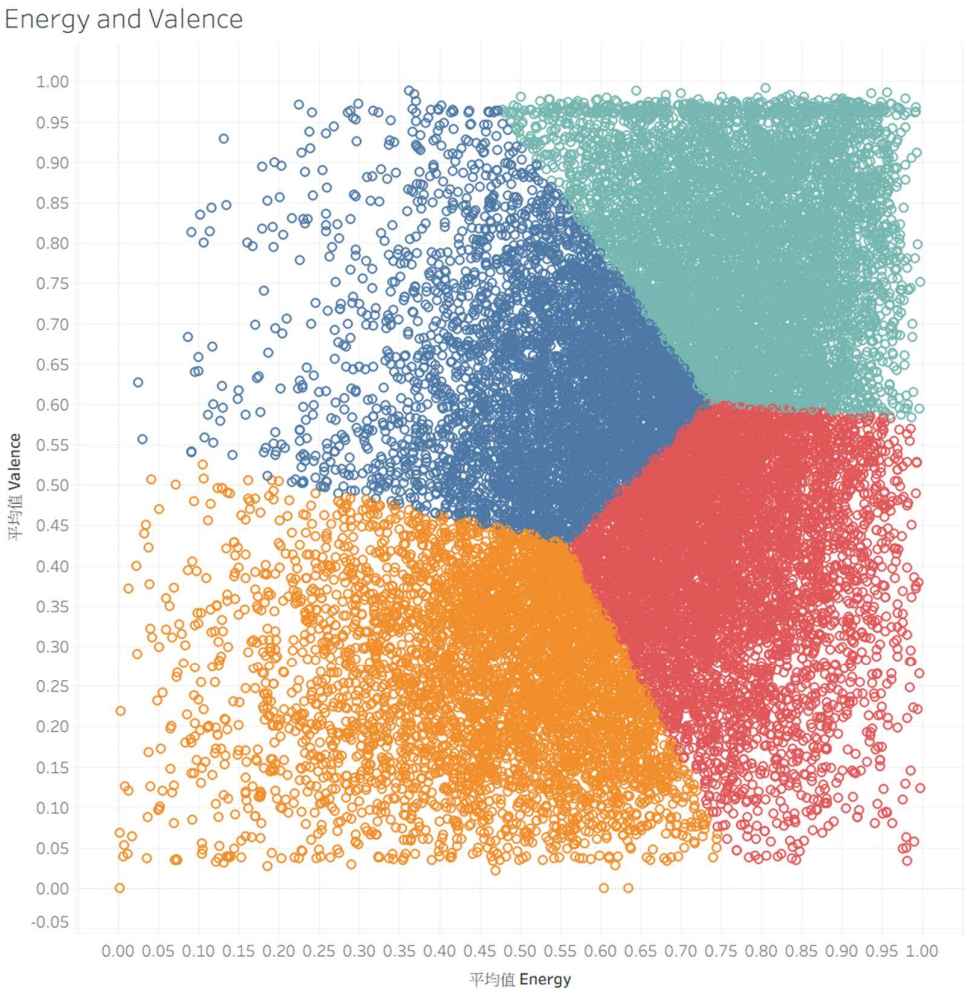


14. The difference music of artists

Artirst Style

Artists	Album Name	
12os Pithikos, Dani Gambino, BoyPan..	Still Standing	words oriented
13 Killoki, Şehinşah	Pembe Yalanlar	song and words balanced
13K, Gleb	BA Connect	words oriented
14 Casper, Bon Nghiêm	SỐ KHÔNG	song and words balanced
14 Casper, Bon Nghiêm, buitruonglinh	SỐ KHÔNG	pure music
21 Savage	american dream	words oriented
21 Savage, Brent Faiyaz	american dream	word leads music
21 Savage, Burna Boy, Metro Boomin	american dream	word leads music
21 Savage, Doja Cat	american dream	word leads music
21 Savage, Lil Durk, Metro Boomin	american dream	words oriented
21 Savage, Mariah the Scientist	american dream	words oriented
21 Savage, Mikky Ekko, Tommy Newp..	american dream	word leads music
21 Savage, Summer Walker	american dream	words oriented
21 Savage, Travis Scott, Metro Boomin	american dream	words oriented
21 Savage, Young Thug, Metro Boomin	american dream	word leads music
23	Bland Sagor & Va..	words oriented
	Förstår om du in..	words oriented
	HOLD UP	words oriented
	Lucky Luciano	words oriented
	PSG	song and words balanced
23, 01an	Förstår om du in..	words oriented
	Kontroversiell	words oriented
	Kontroversiell (I..	words oriented
23, C.Gambino	M.O.B	words oriented
23, Mala	Förstår om du in..	words oriented
23, Roc Boyz	Feeling Myself (..	words oriented
23, Shenzi Beats	Drama	words oriented
	En gång till	words oriented
	Kod Röd	words oriented
	Säg Mig	words oriented
24K, Asme, ROMANOS	Med oss	song and words balanced
25, ADAAM, Kayen	BONANZA	words oriented
25, Greekazo, Sweyway	BUNTAR O SÄNT	word leads music

15. The clustering of Music energy and valence.



Advanced technique task

This task is a freestyle from your idea. First, you have to create dashboards for your project purpose. Bear in mind that you must complete one of the dashboards by using visualized plottings from the last section and drilling some valuable findings. Furthermore, you still have to plot some graphs not mentioned in the previous section and consolidate them as a story-telling dashboard. The below is a dashboard for your reference. You may follow the guidelines of Steve Few informative dashboards for designing a good dashboard. Remember, a good dashboard is not an ornament; it should be a tool for assisting your findings. PLEASE DON'T USE A SINGLE TYPE OF GRAPH FOR YOUR DASHBOARD.