Final Project: Exploratory Data Analysis with SQL

Talib Izhar · Follow
4 min read · Jun 7

Listen Share

Introduction

This project is part of the course SQL: A Practical Introduction for Querying Databases. Here is the source. I will be answering the questions asked on MySQL.

Project Overview

Imagine you have been hired by a non-profit organization that strives to improve socio-economic conditions and educational outcomes for children and youth in the City of Chicago. Your job is to analyze the census, crime, and school data.

You will be asked questions that will help you understand the data just like a real world data professional would. You will be assessed on the correctness of both your SQL queries and results.

Task I: Review and familiarize yourself with the datasets

To complete the assignment problems you will be using three datasets that are available on the city of Chicago' s Data Portal:

1. Socioeconomic Indicators in Chicago
2. Chicago Public Schools
3. Chicago Crime Data

1. Socioeconomic Indicators in Chicago

This dataset contains a selection of six socioeconomic indicators of public health significance and a "hardship index," for each Chicago community area, for the years 2008 – 2012.

A detailed description of this dataset and the original dataset can be obtained from the Chicago Data Portal at: https://data.cityofchicago.org/Health-Human-Services/Census-Data-Selected-socioeconomic-indicators-in-C/kn9c-c2s2

2. Chicago Public Schools

This dataset shows all school level performance data used to create CPS School Report Cards for the 2011-2012 school year.

A detailed description of this dataset and the original dataset can be obtained from the Chicago Data Portal at: https://data.cityofchicago.org/Education/Chicago-Public-Schools-Progress-Report-Cards-2011-/9xs2-f89t

3. Chicago Crime Data

This dataset reflects reported incidents of crime (with the exception of murders where data exists for each victim) that occurred in the City of Chicago from 2001 to present, minus the most recent seven days.

A detailed description of this dataset and the original dataset can be obtained from the Chicago Data Portal at: https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2

> *This dataset is quite large - over 1.5GB in size with over 6.5 million rows. For the purposes of this assignment we will use a much smaller subset of this dataset.*

fig.1 Snapshot of Task 1

Task II: Load the datasets in database tables

I downloaded the datasets available in '.sql ' format and imported all the tables one by one in MySQL database 'practicedb' .

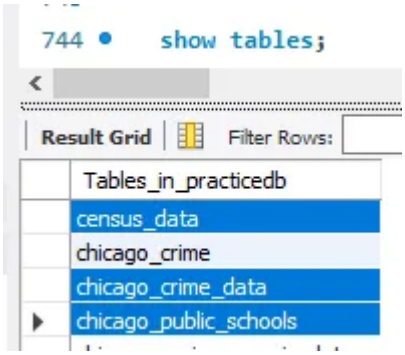


fig2 Imported tables

Task III: Write and execute queries to analyze the data

> *Problem 1*
>
> *Find the total number of crimes recorded in the CRIME table.*

Each case number is for a particular crime so counting distinct case numbers will answer the question.

```
SELECT COUNT(DISTINCT case_number) AS total_crime FROM chicago_crime_data;
```

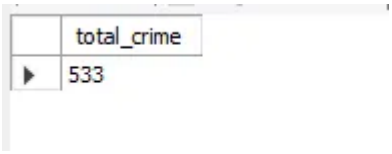


Fig.3 Result of above query

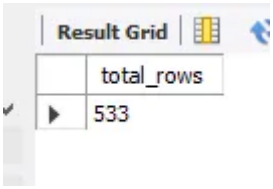
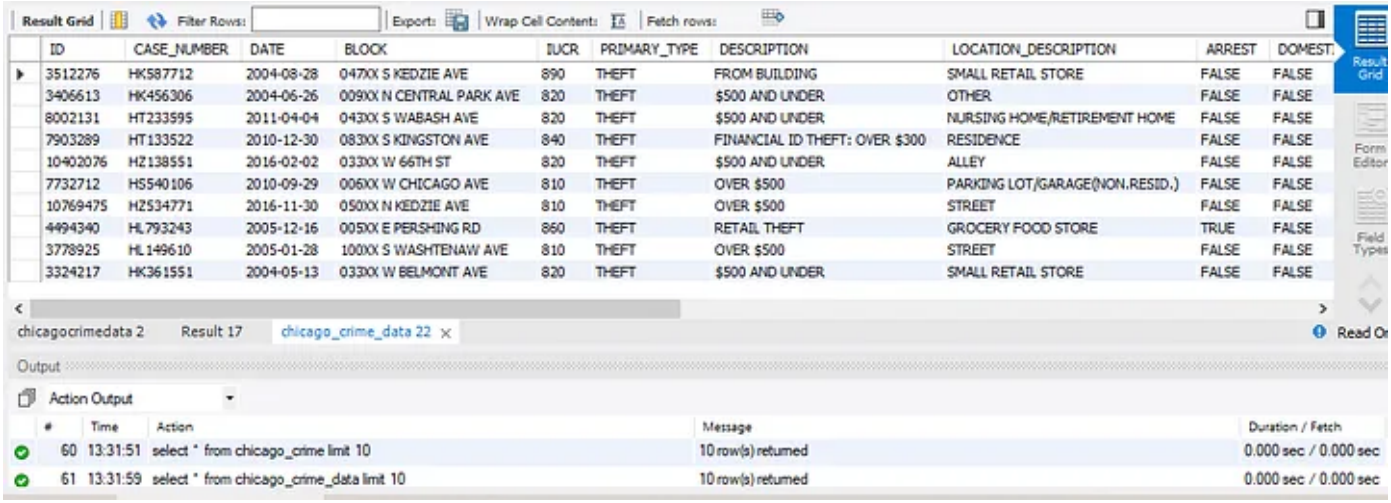


Fig.4

> *Problem 2*
>
> *Retrieve first 10 rows from the CRIME table.*

```
SELECT* FROM chicago_crime_data LIMIT 10;
```



ID	CASE_NUMBER	DATE	BLOCK	IUCR	PRIMARY_TYPE	DESCRIPTION	LOCATION_DESCRIPTION	ARREST	DOMEST...
3512276	HK587712	2004-08-28	047XX S KEDZIE AVE	890	THEFT	FROM BUILDING	SMALL RETAIL STORE	FALSE	FALSE
3406613	HK456306	2004-06-26	009XX N CENTRAL PARK AVE	820	THEFT	$500 AND UNDER	OTHER	FALSE	FALSE
8002131	HT233595	2011-04-04	043XX S WABASH AVE	820	THEFT	$500 AND UNDER	NURSING HOME/RETIREMENT HOME	FALSE	FALSE
7903289	HT133522	2010-12-30	083XX S KINGSTON AVE	840	THEFT	FINANCIAL ID THEFT: OVER $300	RESIDENCE	FALSE	FALSE
10402076	HZ138551	2016-02-02	033XX W 66TH ST	820	THEFT	$500 AND UNDER	ALLEY	FALSE	FALSE
7732712	HS540106	2010-09-29	006XX W CHICAGO AVE	810	THEFT	OVER $500	PARKING LOT/GARAGE(NON.RESID.)	FALSE	FALSE
10769475	HZ534771	2016-11-30	050XX N KEDZIE AVE	810	THEFT	OVER $500	STREET	FALSE	FALSE
4494340	HL793243	2005-12-16	005XX E PERSHING RD	860	THEFT	RETAIL THEFT	GROCERY FOOD STORE	TRUE	FALSE
3778925	HL149610	2005-01-28	100XX S WASHTENAW AVE	810	THEFT	OVER $500	STREET	FALSE	FALSE
3324217	HK361551	2004-05-13	033XX W BELMONT AVE	820	THEFT	$500 AND UNDER	SMALL RETAIL STORE	FALSE	FALSE

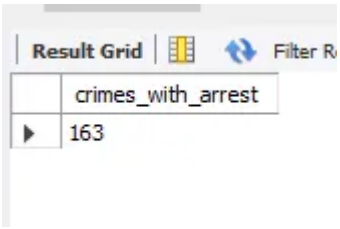
Fig.5 Result of query

Problem 3

How many crimes involve an arrest?

Arrest column is a boolean type either TRUE or FALSE.

```
SELECT COUNT(*) AS crimes_with_arrest
FROM chicago_crime_data
WHERE arrest = 'TRUE';
```



crimes_with_arrest
163

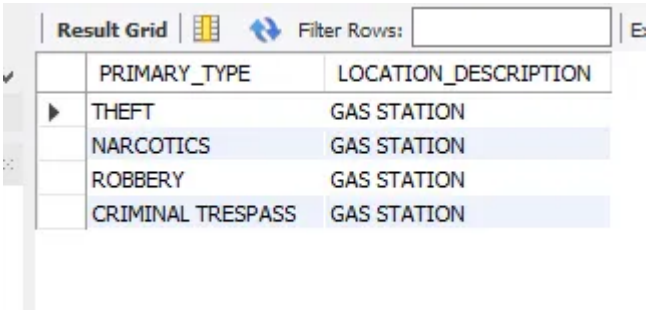
Fig.6 Result of above query

Problem 4

Which unique types of crimes have been recorded at GAS STATION locations?

Crime type is in PRIMARY_TYPE field and location is in LOCATION DESCRIPTION field.

```
SELECT DISTINCT PRIMARY_TYPE, LOCATION_DESCRIPTION
FROM chicago_crime_data
WHERE LOCATION_DESCRIPTION
LIKE '%Gas station%';
```



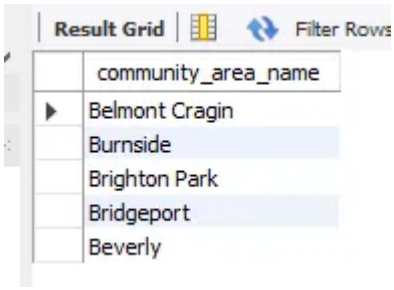
Result Grid | Filter Rows:

PRIMARY_TYPE	LOCATION_DESCRIPTION
THEFT	GAS STATION
NARCOTICS	GAS STATION
ROBBERY	GAS STATION
CRIMINAL TRESPASS	GAS STATION

Fig. 7 Result of query

> *Problem 5*
>
> *In the CENUS_DATA table list all Community Areas whose names start with the letter 'B'.*

```
SELECT community_area_name
FROM CENSUS_DATA
WHERE community_area_name
LIKE 'b%';
```



Result Grid | Filter Rows

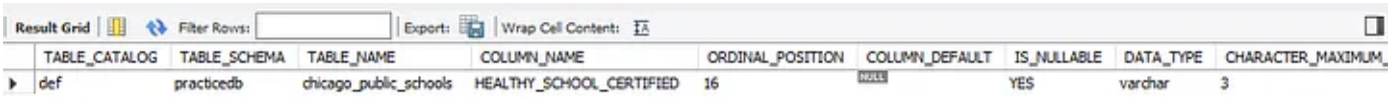
community_area_name
Belmont Cragin
Burnside
Brighton Park
Bridgeport
Beverly

Fig.8 Result of above query

> *Problem 6*
>
> *Which schools in Community Areas 10 to 15 are healthy school certified?*

Sometimes it's hard to find the exact column names and which table have this column, so we can find the column names and the table with query like below:

```
SELECT * FROM INFORMATION_SCHEMA.COLUMNS
WHERE COLUMN_NAME LIKE '%healthy%'
ORDER BY TABLE_NAME ;
```

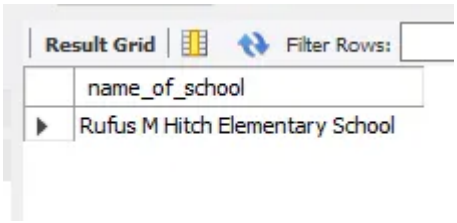


Result Grid | Filter Rows: | Export: | Wrap Cell Content:

TABLE_CATALOG	TABLE_SCHEMA	TABLE_NAME	COLUMN_NAME	ORDINAL_POSITION	COLUMN_DEFAULT	IS_NULLABLE	DATA_TYPE	CHARACTER_MAXIMUM_
def	practicedb	chicago_public_schools	HEALTHY_SCHOOL_CERTIFIED	16	NULL	YES	varchar	3

Fig.9 Result of query

```
SELECT name_of_school
FROM CHICAGO_PUBLIC_SCHOOLS
WHERE Healthy_School_Certified = 'Yes'
AND Community_Area_Number
BETWEEN 10 AND 15;
```



Result Grid | Filter Rows:

name_of_school
Rufus M Hitch Elementary School

FIg. 10 Only one school

> *Problem 7*
>
> *What is the average school Safety Score?*

```
/*Rounded the figure to 2 decimals.*/
SELECT ROUND(AVG(Safety_Score),2)AS avg_sfety_score
FROM CHICAGO_PUBLIC_SCHOOLS ;
```

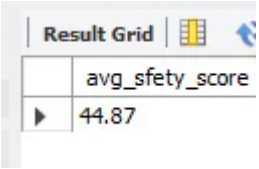


Fig. 11

> *Problem 8*
>
> *List the top 5 Community Areas by average College Enrollment [number of students]*

```
SELECT  Community_Area_Name, AVG(College_Enrollment) AS AVG_ENROLLMENT
FROM CHICAGO_PUBLIC_SCHOOLS
GROUP BY Community_Area_Name
ORDER BY AVG_ENROLLMENT DESC
LIMIT 5;
```

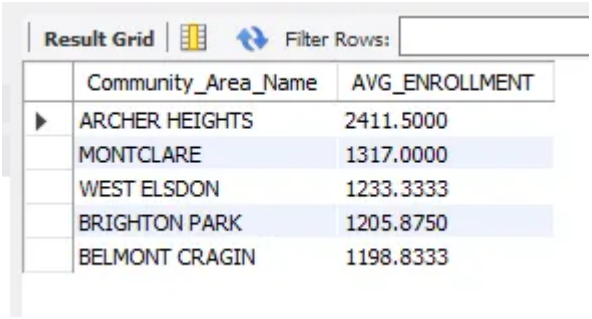


FIg. 12 Result of query

> *Problem 9*
>
> *Use a sub-query to determine which Community Area has the least value for school Safety Score?*

Safety score is in character data type in table, so first I found distinct safety scores.

```
SELECT DISTINCT safety_score
FROM chicago_public_schools
ORDER BY safety_score ;
```

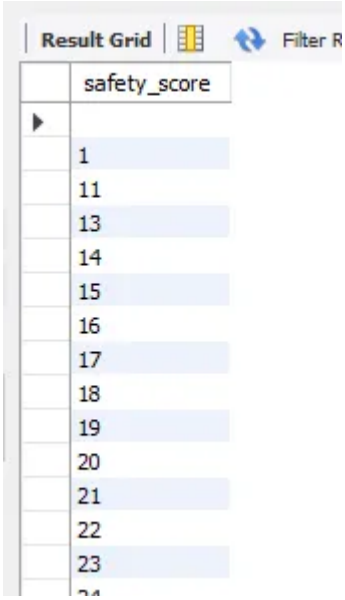


Fig. 13

Blanks are also present, so the least value is 1.

```
SELECT COMMUNITY_AREA_NAME, safety_score
FROM (SELECT COMMUNITY_AREA_NAME, safety_score
FROM chicago_public_schools WHERE safety_score = 1) school;
```

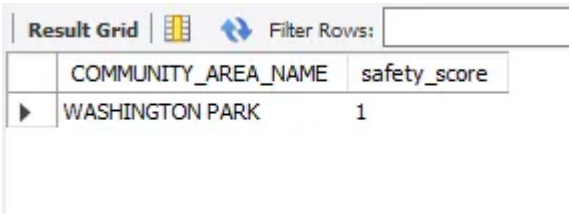
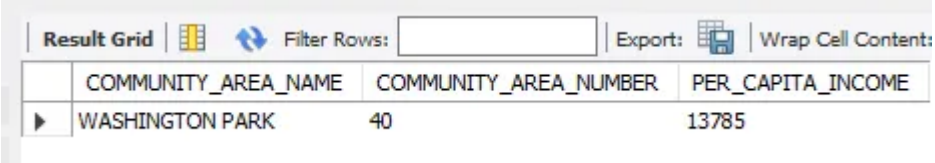


Fig.14 Only one area

> *Problem 10*
>
> *[Without using an explicit JOIN operator] Find the Per Capita Income of the Community Area which has a school Safety Score of 1.*

```
SELECT cs.COMMUNITY_AREA_NAME, cs.COMMUNITY_AREA_NUMBER, PER_CAPITA_INCOME
FROM census_data cd, chicago_public_schools cs
```

```
WHERE cs.COMMUNITY_AREA_NUMBER=cd.COMMUNITY_AREA_NUMBER AND safety_score=1;
```



Result Grid | Filter Rows: | Export: | Wrap Cell Content:

	COMMUNITY_AREA_NAME	COMMUNITY_AREA_NUMBER	PER_CAPITA_INCOME
▶	WASHINGTON PARK	40	13785

Fig. 15 Result of above query

> *Thank you for reading! If you want to give any suggestion or any feedback please feel free to comment (:*

Exploratory Data Analysis | MySQL | Data Analytics

Follow

Written by Talib Izhar

4 Followers

EXCEL | SQL | Power BI | R

More from Talib Izhar

ment problems you will be using three datasets that are available on the city of Chicago' s Data Portal:

licators in Chicago

ools

a

mic Indicators in Chicago

selection of six socioeconomic indicators of public health significance and a "hardship index," for each Chicago community area, for the years 2008 – 2012.

of this dataset and the original dataset can be obtained from the Chicago Data Portal at: https://data.cityofchicago.org/Health-Human-Services/Census-Data-Selected-socioeconomic-indi

blic Schools

chool level performance data used to create CPS School Report Cards for the 2011-2012 school year.

of this dataset and the original dataset can be obtained from the Chicago Data Portal at: https://data.cityofchicago.org/Education/Chicago-Public-Schools-Progress-Report-Cards-2011-/9

me Data

orted incidents of crime (with the exception of murders where data exists for each victim) that occurred in the City of Chicago from 2001 to present, minus the most recent seven days.

of this dataset and the original dataset can be obtained from the Chicago Data Portal at: https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2

e large - over 1.5GB in size with over 6.5 million rows. For the purposes of this assignment we will use a much smaller subset of this dataset.

Talib Izhar

Final Project: Advanced SQL Techniques

Answering the questions of project

6 min read · Jun 10


1

Talib Izhar

PwC Switzerland Power BI Virtual Case Experience

Sharing my work under the program and experience in Pwc Power BI virtual Internship on Forage.

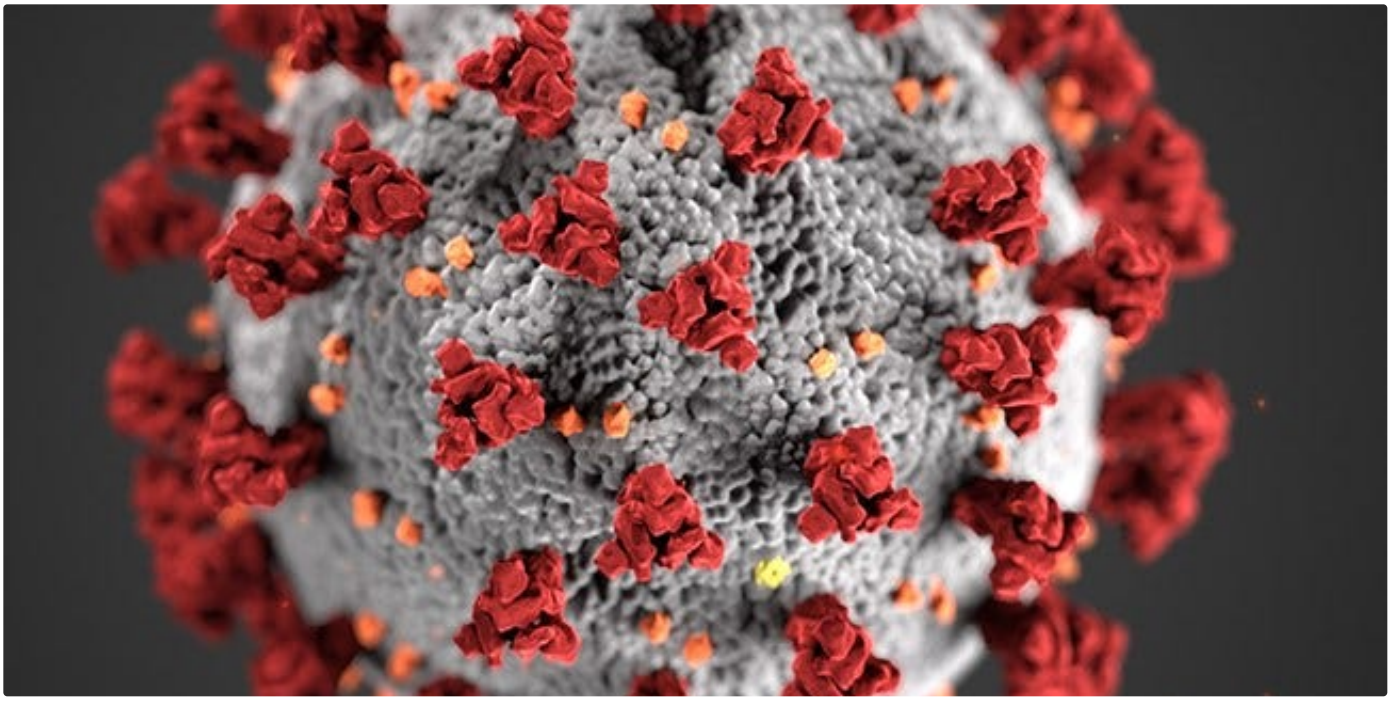
5 min read · Jul 26

Talib Izhar

Power BI Olist Dashboard

Interactive KPI dashboard

4 min read · Jun 27



Talib Izhar

Visualizing Covid-19

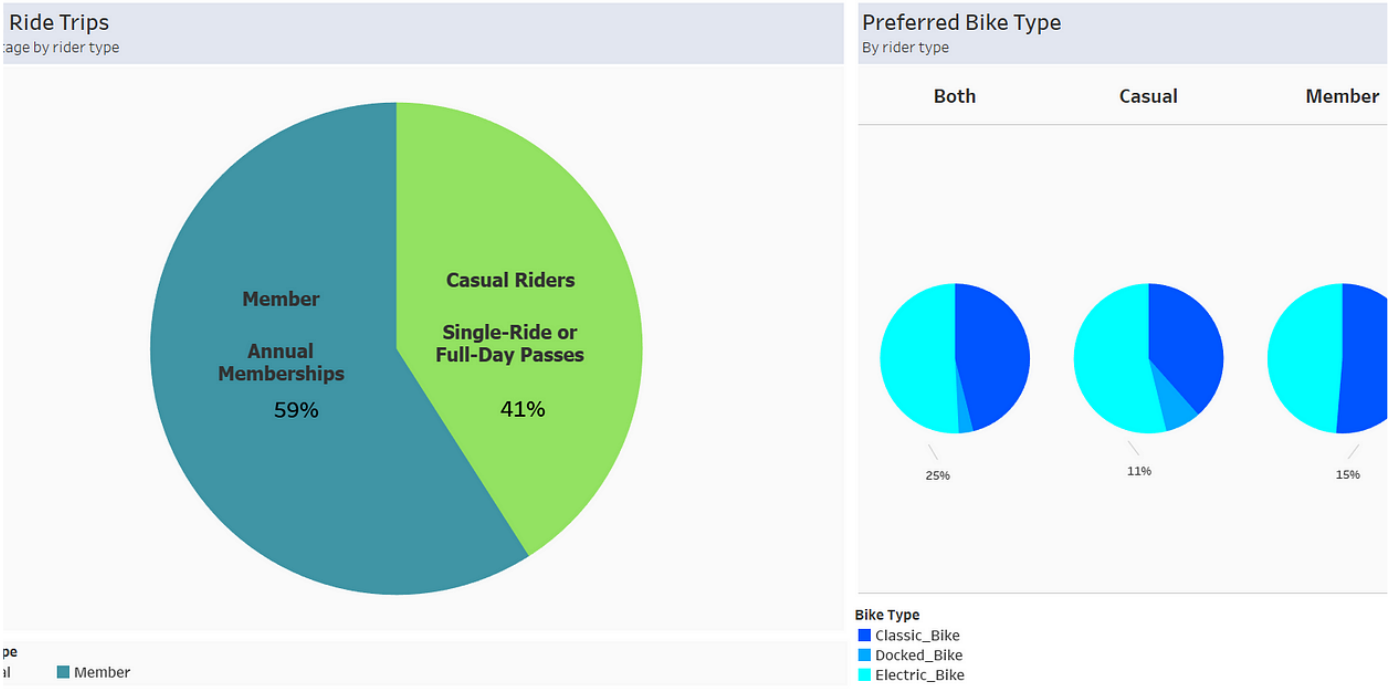
Unfolding the events during COVID-19 and finding the hardest-hit country till March 2020.

7 min read · Aug 27

2

See all from Talib Izhar

Recommended from Medium



Richmund Garces Allorde

Cyclistic Bike-Share Analysis using Excel, SQL, & Tableau

Introduction

4 min read · Jul 7

5

Mohamed Abdelaal elesely

10 Data Analysis Projects to land your first Job.

To land a Job in a very technical field such as Data analysis you need to have a solid portfolio, This portfolio will be your identity Card...

4 min read · May 5

672 8

Lists

Practical Guides to Machine Learning

10 stories · 380 saves

New_Reading_List

174 stories · 92 saves

Modern Marketing

33 stories · 106 saves

Now in AI: Handpicked by Better Programming

266 stories · 119 saves

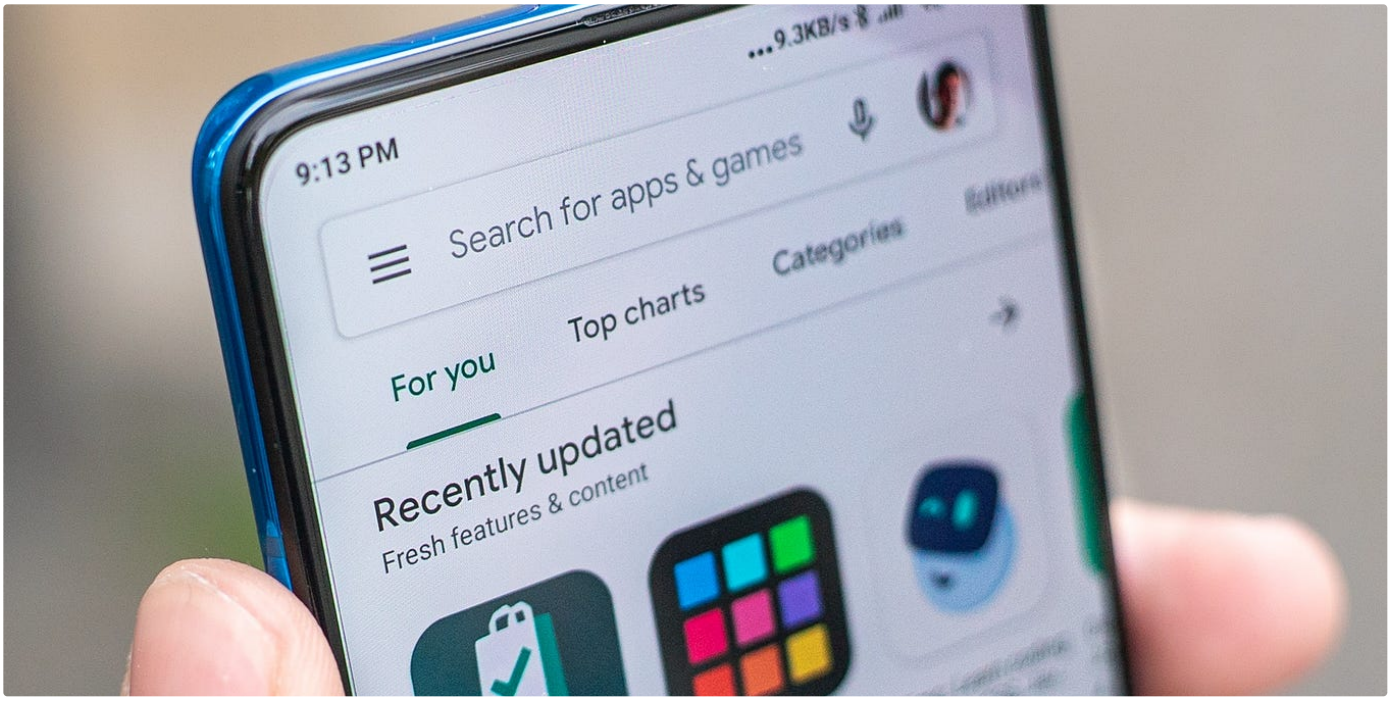
Adeola

Google Data Analytics Capstone Project: Bellabeat Case study

This is a project documentation for Bellabeat case study from the Google Data Analytics Course. The analysis follows the 6 steps of Data...

13 min read · Jul 12

96



Asheley Mudzingwa

Exploratory Data Analysis (EDA) on Google Play Store Dataset

Exploring App Trends: Analyzing Categories, Ratings, and Downloads

3 min read · Aug 27

55

Iza Stań

Silent Heros of Analytics: Data Preprocessing 101—Data Cleaning

We have all heard the age-old saying, “Garbage in, garbage out.” It unveils a truth in data analytics: the quality of your insights comes...

9 min read · Aug 26

30 1


Sodiq Babatunde

My First Three Months Experience As a Data Analyst

Hmmmmmm…. Where do I start from???. Okay, I resumed work as a data analyst on the 9th of January, 2023. Yeah!!! It's quite an exciting…

3 min read · Apr 17


45


4

See more recommendations