===================================================================

BUSINESS ANALYSIS & REPORTING - FINAL PROJECT

Student Name: Trang Bui

StudentID: W0753523

===================================================================

**PART 1**

**1. Write a query to display the top 10 hotspots for coronavirus (COVID – 19) cases in the**

**USA by County showing confirmed cases, population and the median age. Create the**

**appropriate chart and format properly with Chart Title and Axis Title. Also comment on**

**the chart in MS Word file.**

/\*EXPLANATION

For visualization purpose, I decided to add 2 more columns to show portion of confirmed cases out of state population, and the remain portion which has no confirmation of virus. The reason is there is a very big gap of the number among population, confirmation cases, and median age.

I decided to calculate average of median age because we have data of median age by zip\_code. There are many zipcode in 1 county.

\*/

select top 10 t.Admin2 CountyName, ConfirmedCases, zct.CountyPop CountyPopulation, avg(zcs.MedianAge) CountyAvgMedianAge,

(cast(ConfirmedCases as float)\*100/zct.CountyPop) as Pct\_ConfirmedCasesOutofPopulation,

(100-(cast(ConfirmedCases as float)\*100/zct.CountyPop)) as pct\_NoConfirm\_Population

from [dbo].[ZipCounty] zct

inner join [dbo].[ZipCensus] zcs on zct.ZipCode=zcs.zcta5

inner join [dbo].[time\_series\_covid\_19\_confirmed\_US] t on zct.CountyFIPS=t.FIPS

group by t.Admin2, ConfirmedCases, zct.CountyPop

order by ConfirmedCases desc;

“The combo chart presents statistics data including the percentage of virus confirmation cases out of population, population, median age in top 10 states having the biggest number of cases in USA. The stacked columns address percentage of virus cases out of population. It can be easily found that the rate confirmed cases in New York reached the highest rate at 15% out of its population. Whereas the other counties have the rate of covid19 cases lower than 5 percent of their population. Although the percentage is low, the higher number of populations, the bigger number of confirmed cases. Median age in these countries are from 33-43 that told that they have young population.”

**2. Write a query to display male and female deaths by Country who have never visited Wuhan and have been asymptomatic (showing no symptoms). Order Deaths in descending order. Create a stacked bar chart. Also comment on the chart in MS Word file.**

select Country,

count(CASE WHEN Gender = 'Male' THEN 1 END) as Male,

count(CASE WHEN Gender = 'FeMale' THEN 1 END) as Female,

count(CASE WHEN Gender NOT IN ('Male','FeMale') THEN 1 END) as Unknow\_Gender,

count(\*) as Total\_Deaths

from dbo.COVID19\_line\_list\_data cv

where death=1 and visiting\_Wuhan=0 and symptom is null

group by country

order by Total\_Deaths desc;

“The given stacked column chart presents the number of deaths by gender due to Covi19, who have never visited Wuhan. The highest stacked column showed the biggest number of deaths are in China with 27 males and 12 females. This proved that the pandemic spreaded easily internal country. France where is in different continent and very far away from Wuhan, got 2 case of deaths for male only which is the lowest number in graph. South Korea and Iran have deaths at 9 and 4 respectively. Gender of death is not definite in Iran while in France 8 males and 1 female are dead.”

**3. Write a query showing an age distribution chart among cases of COVI9-19 patients. Group**

**by categories below. What % group of the population has highest admission. Order from**

**highest to lowest.**

**o Baby Boomers (Roughly 50+ years old)**

**o Generation X (Roughly 35 – 50 years old)**

**o Millennials, or Generation Y (18 – 34 years old)**

**o Generation Z, or generation (17 & younger)**

**Draw appropriate graph/chart in excel. Also comment on the chart in MS Word file.**

select Age\_Group, Number\_of\_Patients, cast(Number\_of\_Patients as float)/(sum(Number\_of\_Patients) over()) \* 100 as pct\_Patients

from (

select case when age>50 then 'Baby Boomers'

when age between 35 and 50 then 'Generation X'

when age between 18 and 34 then 'Generation Y'

when age <=17 then 'Generation Z'

end as Age\_Group,

count(\*) Number\_of\_Patients

from dbo.COVID19\_line\_list\_data

group by case when age>50 then 'Baby Boomers'

when age between 35 and 50 then 'Generation X'

when age between 18 and 34 then 'Generation Y'

when age <=17 then 'Generation Z'

end

) pt

order by Number\_of\_Patients desc;

“The pie chart demonstrated the percentage of deaths by age group which are divided to 4 groups including Baby Boomers for age greater 50, Generation Z for age below 18, Generation Y for age between 18 and 34, and Generation X for age between 35 and 50. It is easily realized the number of deaths of Baby Boomers got the biggest portion of the pie which took 39%. The age group X and Z account nearly same portions which are 22 and 25%. The lowest percentage for age group of generation Y showed that this group has least impact by Covid-19 than other groups.”

**4. Does Family Income impact COVID 19 cases? We want to see if any correlation is present**

**does higher income county show less cases, more cases or equally distributed cases? What**

**about foreign-born residents do they have higher, lower or equally distributed cases of**

**COVID 19 compared to Born in US residents? Foreign Born VS Born in USA. Depict a**

**chart showing both these scenarios and provide a trendline. You can choose to create one**

**query or two separate queries and combine the results in Excel. Comment on the**

**distribution here**.

4a--- We want to see if any correlation is present does higher income county show less cases, more cases or equally distributed cases?

select zc.County, t.ConfirmedCases, avg(zc.MedianEarnings) CountyAvgMedianEarnings

from dbo.time\_series\_covid\_19\_confirmed\_US t

inner join dbo.ZipCensus zc on t.FIPS=zc.Fipco

group by zc.County, t.ConfirmedCases

order by CountyAvgMedianEarnings desc;

“The scatter plot shows the relationship between income and covid-19 impaction. Basing on the trend line, it can be easily recognized that there is no strong correlation existing. The points are not over the linear line. The density of confirmed cases is focused at the range of income from 25,000 to 45,000.”

4b- /\* EXPLANATION

I cannot find out information of place of Birth of Patients which are confirmed cases.

Therefore, I assume that "COVID19\_line\_list\_data.country" is the place of born.

This means that a patient was born in USA, the country field value is "USA", other values are foreign born.

COVID19\_line\_list\_data.location contents information of both country name and state names, it is not mapped to county.

So, I cannot plot data for relationship b/w confirmed cases of USA Born and Foreign Born. Just visualizing for the gap b/w these number of cases.

\*/

select count(case when country = 'USA' then 1 end) as Nb\_Cases\_USABorn

,count(case when country <> 'USA' then 1 end) as Nb\_Cases\_ForeignBorn

from COVID19\_line\_list\_data

“The clustered bar chart so the gap of the number of confirmed cases of USA Born patients and Foreign Born. There is a very big distance between these numbers. The number of cases of foreign born is 1066 while there is only 18 cases for USA Born. The number of USA Born accounts 2% out of total only.”

**5. Currently in the news they are saying African Americans are dying at a higher rate. Use**

**the data provided to support or refute this claim showing demographics among all**

**ethnicities in US. Create the appropriate chart(s) and format properly with Chart Title and**

**Axis Title. Also comment on the chart in MS Word file**.

/\*EXPLANATION

I couldn't find information of African Americans' deaths. So, I decided to summary number of deaths due to Covid-19 over the world in table COVID19\_line\_list\_data where there is no data mentioning in the news to decline to it

\*/

select country, count(\*) as number\_of\_death\_by\_country

from dbo.COVID19\_line\_list\_data

where death>0

group by country

order by number\_of\_death\_by\_country desc;

“The column chart shows the number of deaths by countries where there is no data of African Americans deaths. **Therefore, the news that African Americans are dying at a higher rate is declined**. Data of “COVID19\_line\_list\_data” is selected until Feb-28th,2020 only”

**6. How many confirmed cases were reported worldwide each week? Create the appropriate**

**chart(s) and format properly with Chart Title and Axis Title. Also comment on the chart in**

**MS Word file.**

select datepart(year,reporting\_date) year, datepart(wk,reporting\_date) weekinyear, count(\*) nbReportedCases

from dbo.COVID19\_line\_list\_data

where datepart(year,reporting\_date) != 1900

group by datepart(year,reporting\_date), datepart(wk,reporting\_date)

order by year, weekinyear;

“The line chart demonstrates the trend of weekly reported cases. In general, the trend line shows increasing of number of cases reported. Starting from the third week in year with a lowest number at 3. The number of cases raised up strongly after 1 week only that reached 199 cases. The pandemic temperature seemed to decrease week by week until the seventh. However, once more times the reported cases number rose again through the week eighth and climbed up reach the peak at the nineth week in 2020 at 274 cases.”

**7. How many confirmed cases were reported country wide each day? Create the appropriate**

**chart(s) for top 10 countries and format properly with Chart Title and Axis Title. Also**

**comment on the chart in MS Word file**.

--7a. How many confirmed cases were reported country wide each day?

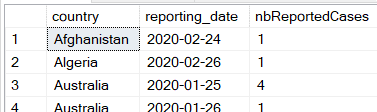
select country, reporting\_date, count(1) nbReportedCases

from dbo.COVID19\_line\_list\_data

where datepart(year,reporting\_date) != 1900

group by country, reporting\_date

order by country, reporting\_date;



--7b. -- Create the appropriate chart(s) for top 10 countries and format properly with Chart Title and Axis Title

with top10countries as

(

select top 10 country, count(1) nbReportedCases

from dbo.COVID19\_line\_list\_data

where datepart(year,reporting\_date) != 1900

group by country

order by nbReportedCases desc

)

select country, reporting\_date, count(1) nbReportedCases

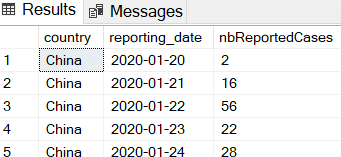
from dbo.COVID19\_line\_list\_data cv

where datepart(year,reporting\_date) != 1900

and exists (select 1 from top10countries t10 where t10.country = cv.country)

group by country, reporting\_date

order by country, reporting\_date;



“The area chart demonstrates the trend of daily reported cases in top 10 countries where had the biggest number of positive cases confirmed. Since Covid-19 was found from China, the number of new cases were very high every day in this country on Jan-2020. The pandemic quickly spread to other countries such as Thailand, Singapore, France, South Korea… with the daily case below 5. The number of new cases increased strongly from mid of Feb that was a sign of pandemic outbreak, especially in South Korea where the daily new cases rose up to 55, and other countries including Japan, Spain had daily new cases from 10 to 20.”

**8. Find Top Ten Closest Zip Codes to the US Geographic Center (Latitude:39.80 and**

**Longitude: -98.60 ) with COVID-19 confirmed cases. No Excel chart required here SQL**

**only**.

-- use SRID=4326 to convert distance to metre

SELECT \*

FROM sys.spatial\_reference\_systems

WHERE spatial\_reference\_id = 4326;

-- Find Top Ten Closest Zip Codes to the US Geographic Center (Latitude:39.80 and Longitude: -98.60 ) with COVID-19 confirmed cases

Select top 10 ZipCode, POName, CountyName, State,

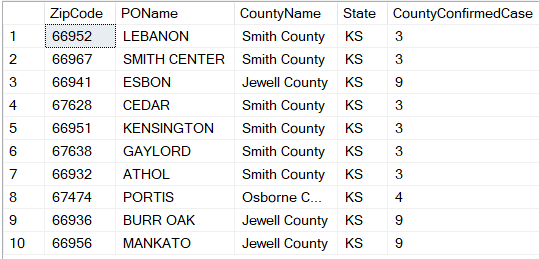
(select sum(ConfirmedCases) CountyConfirmedCases

from dbo.time\_series\_covid\_19\_confirmed\_US t

where zc.CountyFIPS=t.FIPS) CountyConfirmedCase

from dbo.ZipCounty zc ---time\_series\_covid\_19\_confirmed\_US

order by GEOGRAPHY::Point(39.80, -98.60, 4326).STDistance(GEOGRAPHY::Point(Latitude, Longitude, 4326));



**9. Find the difference (No. of days) between symptom\_onset and reporting date. Then write another query to find No. of COVID-19 cases for each difference (No. of Days). No Excel chart required. SQL only.**

-- Find the difference (No. of days) between symptom\_onset and reporting date.

select datediff(day,symptom\_onset,reporting\_date) dayDiffSymptomReporting,

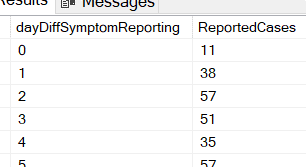
count(\*) ReportedCases

from dbo.COVID19\_line\_list\_data

where datepart(year,reporting\_date) != 1900

and datepart(year,symptom\_onset) != 1900

group by datediff(day,symptom\_onset,reporting\_date)



**10. How many unique symptoms of COVID-19 are there in the data set (symptoms can be only ONE or combination of more than one as a single TEXT/STRING as given in the**

**symptoms column). List down No. of cases by gender for each symptom type. Compare**

**this result using suitable chart/graph in Excel.**

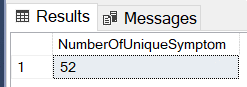
--10a. How many unique symptoms of COVID-19 are there in the data set

[Answer] 52

select count(distinct trim(value)) as NumberOfUniqueSymptom

from dbo.covid19\_line\_list\_data

cross apply string\_split(symptom, ',') s



--10b. List down No. of cases by gender for each symptom type. Compare this result using suitable chart/graph in Excel.

select trim(value) as symptom,

count(CASE WHEN Gender = 'Male' THEN 1 END) as Male,

count(CASE WHEN Gender = 'FeMale' THEN 1 END) as Female,

count(CASE WHEN Gender NOT IN ('Male','FeMale') THEN 1 END) as Unknow\_Gender,

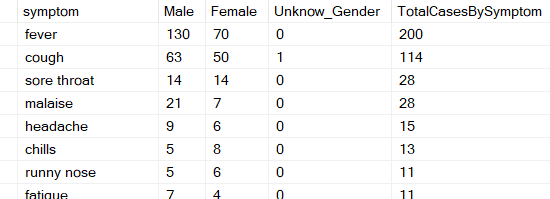
count(\*) TotalCasesBySymptom

from dbo.covid19\_line\_list\_data

cross apply string\_split(symptom, ',') s

group by trim(value)

order by TotalCasesBySymptom desc;



“The stacked column chart compares the number of cases across the list symptoms of Covid-19 and portion of gender in each symptom. Fever is the most common symptom of covid-19 patients both for male and female that occurred on 200 cases in which 130 cases are for male, and 70 is for females. The second popular sign is cough found from 114 cases. Portion for male and female with cough symptom is not much different from gender that is 63 and 50 for male and female respectively. Besides, there are many other symptoms such as malaise, sore throat, headache, chills, runny nose… which met on Covid-19 cases. Most of symptoms exposed clearer for males than females.”

===================================================================

BUSINESS ANALYSIS & REPORTING - FINAL PROJECT

Student Name: Trang Bui

StudentID: W0753523

===================================================================

**PART 2**

1. **What is survival analysis? Explain clearly with an example.**

Survival Analysis is the set of statistical method that are used to analyze time-to-event data. The response could be failure time, survival time, or event time. It is used to estimate the lifespan of a specific population.

For instance, we analyze the average person lives from cancer treatment until death event. We will collect information including living days, patients’ personal factors like disease, cancer, or other treatment, diagnosis, gender, age, etc. With this information, we build a survival model to split them into multiple groups and compare. It will give us the prediction of survival time for new patients and will help us to find the best treatment for them.

Another example, we can use survival analysis to determine the probability of failure of factory machines based on hours of operations and operating conditions.

1. **Compare Survival and Retention**.

Survival is to calculate data from beginning up to a point of time.

Retention is to calculate data in time interval (time period).

1. **Explain hazard probability with an example**.

It uses to assess risk, it calculates the probability of life failure or not survive in a time interval, it’s opposite with the retention rate.

For example, in a model to calculate a person’s number of years left to live, we compare the probability of dying at age 75 of 2 people, one is 70 and one is 30. The person aged 70 will have a higher probability of dying (hazard probability) than the person aged 30 because he still has many time (years) left in his life.

1. **Discuss different types of purchase patterns**

The patterns of all unique set of characterization of all individual customers can be grouped according to:

* Place of purchase: In most cases, customers will purchase items in a number of stores even all items they need in the same store if they afford to do that, unless this is only store they have to access. They don’t remain loyal in one store and visit various stores in different places for price comparisons or brand preferences. So, location must be a consideration for distributors.
* Types and quantity of item purchased: Most of customers will buy their necessary items and there are fewer customers will purchase luxury products. Knowledge of all factors of each item purchased such as price, unit purchasing power, promotion, number of items purchased at a time, etc. are very important for manufacturer, distributor as well as the marketer.
* Time and frequency of purchase: In most cases, customers go shopping according to their feasibility, however not all of them go the same time as each other, so the shops should stay updated with the time and frequency of the purchases. They should also monitor the waiting line at the main door and the checkout counters to optimize their services.
* Method of purchase: This pattern will involve a customer directly walking into a store when it can offer the convenient way to purchase and delivery. Shop that offers purchase method like debit, credit, cash, 2-hour delivery, etc. might influence customer to say yes to purchase items here.

1. **What are the different types of Customer Characteristics that can be investigated in a customer database?**

* Identity data: information of name, person, postal address, telephone, email address, social network, account, job, etc.
* Quantitative data: information of transaction (online/offline), communication channel, online activity, social network activity, customer services, etc.
* Descriptive data: information of family details, lifestyle, career, etc.

1. **How different days of the week impact the revenue of a business.**

Currently, most of companies are using 40-hour work week with 5 working days structure that comes from Henry Ford, the founder of the Ford Motor Company, from 1926. This structure is suitable and adapt for most of industries. For some roles that require more hours, companies use set of shift hours with the same structure, so every employee still get time to refresh during the weekend.

The 4-day or 32-hour work week offers employees flexible hours and a better work-life balance. Employees will have more time for family in day off instead of spending lots of weekend time for grabbing groceries and stuff. The company cost will be reduced while employee don’t come to the office to work for 1 day. The productivity might get influences when working time is less than the common, however it could be not much when the employees are always happier and healthier with the new balance of work and life, so the production still be good.

The 6-day or more work week are still common in Asian countries which most of them are developing countries. This structure helps the revenue of company grow fast, however, it causes unbalanced life for employees and it won’t help them to have a better life with family. This structure should be applied in a short term when it may harm the productivity if the employees are stressful for long time.

1. **Why is month-to-date comparison important?**

Because it helps to compare performance of current month-to-date (MTD) to that of past MTD periods, allows owners, manager or others to get the picture of current trend to take action to bring in extra help if needed.

1. **What is the purpose of datepart() and datename() functions. Give one example for each.**

They return the similar result except the data type, the datepart() returns date part as an integer while the datename() returns the date part as a character string. So, depending on usage purpose, we will use proper function. For example, if we would like to display month name or week day name we should use datename() function. If we want to calculate with the date part or display number of month or week day, we can use the datepart() function.

|  |  |
| --- | --- |
| SQL query | Result |
| DECLARE @dt DATETIME2= current\_timestamp;  SELECT 'year,yyy,yy' date\_part,  DATENAME(year, @dt) resultdatename,  DATEPART(year, @dt) resultdatepart  UNION ALL  SELECT 'quarter, qq, q',  DATENAME(quarter, @dt),  DATEPART(quarter, @dt)  UNION ALL  SELECT 'month, mm, m',  DATENAME(month, @dt),  DATEPART(month, @dt)  UNION ALL  SELECT 'dayofyear, dy, y',  DATENAME(dayofyear, @dt),  DATEPART(dayofyear, @dt)  UNION ALL  SELECT 'day, dd, d',  DATENAME(day, @dt),  DATEPART(day, @dt)  UNION ALL  SELECT 'week, wk, ww',  DATENAME(week, @dt),  DATEPART(week, @dt)  UNION ALL  SELECT 'weekday, dw, w',  DATENAME(weekday, @dt),  DATEPART(weekday, @dt)  UNION ALL  SELECT 'hour, hh' date\_part,  DATENAME(hour, @dt),  DATEPART(hour, @dt)  UNION ALL  SELECT 'minute, mi,n',  DATENAME(minute, @dt),  DATEPART(minute, @dt)  UNION ALL  SELECT 'second, ss, s',  DATENAME(second, @dt),  DATEPART(second, @dt)  UNION ALL  SELECT 'millisecond, ms',  DATENAME(millisecond, @dt),  DATEPART(millisecond, @dt)  UNION ALL  SELECT 'microsecond, mcs',  DATENAME(microsecond, @dt),  DATEPART(microsecond, @dt)  UNION ALL  SELECT 'nanosecond, ns',  DATENAME(nanosecond, @dt),  DATEPART(nanosecond, @dt)  UNION ALL  SELECT 'TZoffset, tz',  DATENAME(tz, @dt),  DATEPART(tz, @dt)  UNION ALL  SELECT 'ISO\_WEEK, ISOWK, ISOWW',  DATENAME(ISO\_WEEK, @dt),  DATEPART(ISO\_WEEK, @dt); | |  |  |  | | --- | --- | --- | | date\_part | resultdatename | resultdatepart | | year,yyy,yy | 2020 | 2020 | | quarter, qq, q | 3 | 3 | | month, mm, m | August | 8 | | dayofyear, dy, y | 222 | 222 | | day, dd, d | 9 | 9 | | week, wk, ww | 33 | 33 | | weekday, dw, w | Sunday | 1 | | hour, hh | 22 | 22 | | minute, mi,n | 19 | 19 | | second, ss, s | 0 | 0 | | millisecond, ms | 163 | 163 | | microsecond, mcs | 163333 | 163333 | | nanosecond, ns | 1.63E+08 | 1.63E+08 | | TZoffset, tz | +00:00 | 0 | | ISO\_WEEK, ISOWK, ISOWW | 32 | 32 | |

1. **Elaborate the purpose of Cast() function with an example.**

The **cast()** function is ANSI standard and it is used to convert a value (from a field, a variable or an expression) of any type into a specific datatype.

For example, there is a task to load a csv file into a table. First step, we load csv file to a temporary table having varchar type for all columns. At second step, we will cleanup data. Next step, we will cast varchar data type into proper data type for all columns.

After the first and second steps, we assume we have data of varchar like this:

create table ztmp\_test(

id varchar(255),

name varchar(255),

numberofitems varchar(255),

created varchar(255)

);

insert into ztmp\_test(id,name,numberofitems,created) values ('1','one','10','8/1/2020');

insert into ztmp\_test(id,name,numberofitems,created) values ('2','two','20','8/2/2020');

And at the third step, we cast data to have proper data type:

create table test(

id int,

[name] varchar(255),

numberofitems int,

created date

);

insert into test(id,name,numberofitems,created)

select cast(id as int), name, cast(numberofitems as int), cast(created as date)

from ztmp\_test;

select \* from test;

1. **Give an example of calculating the extreme rich and extreme poor households.**

According the information of the World Bank, the extreme poor households refer to individuals living in a household with per capita consumption or income below $1.90. See [1] for reference.

According to a 2019 study by economists Emmanuel Saez and Gabriel Zucman found that the average effective tax rate paid by the richest 400 families (0.003%) in the country was 23 percent, a full percentage point lower than the 24.2 percent rate paid by the bottom half of American households. See [2] for reference.

[1]: <https://ourworldindata.org/extreme-poverty-in-rich-countries-what-we-know-and-what-we-dont-know>

[2]: <https://en.wikipedia.org/wiki/Wealth_inequality_in_the_United_States>