

Udacity's Data Analyst Nanodegree Program

PROJECT #1 - EXPLORE WEATHER TRENDS

STUDENT - MEHROL BAZAROV

In my first project in the Udacity's Data Analyst Nanodegree program, I analyzed local and global temperature data and compare the temperature trends where I live to overall global temperature trends. In this report you can find things which are listed below:

OUTLINE:

Extracting data from database	- page 2
Prepared CSV file	- page 3
Open up the CSV	- page 5
Moving Average Calculation	- page 5
Key consideration to choose visualization	- page 6
Line Chart	- page 7
Six Observations	- page 7
<hr/>	
Additional visualization #1 with summary	- page 8
Additional remarkable insight #2	- page 9

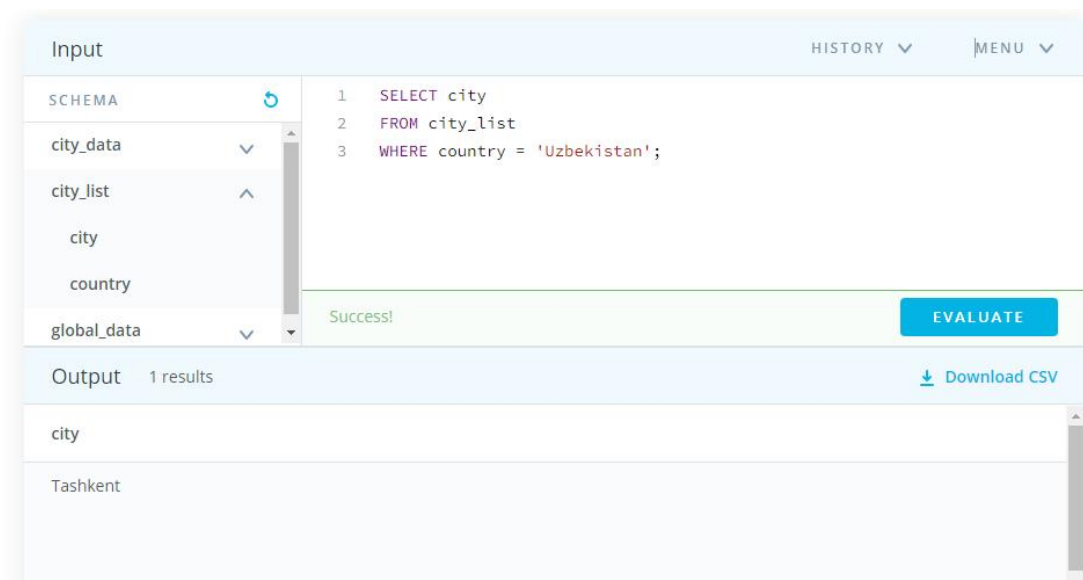
Extracting data from database

Firstly, I used **SQL** to extract data from database to find whether my country is listed in the table called **city_list** which contains two attributes *city* and *country*.

```
SELECT city
```

```
FROM city_list
```

```
WHERE country = 'Uzbekistan';
```



Output of this query is only the Tashkent city where I live. It is good. Students who are living in Uzbekistan but their city is not Tashkent will choose Tashkent for their analysis. So they have only one option. They may not really get accurate results. Why I am telling this because, I found out that some countries contains lots of cities:

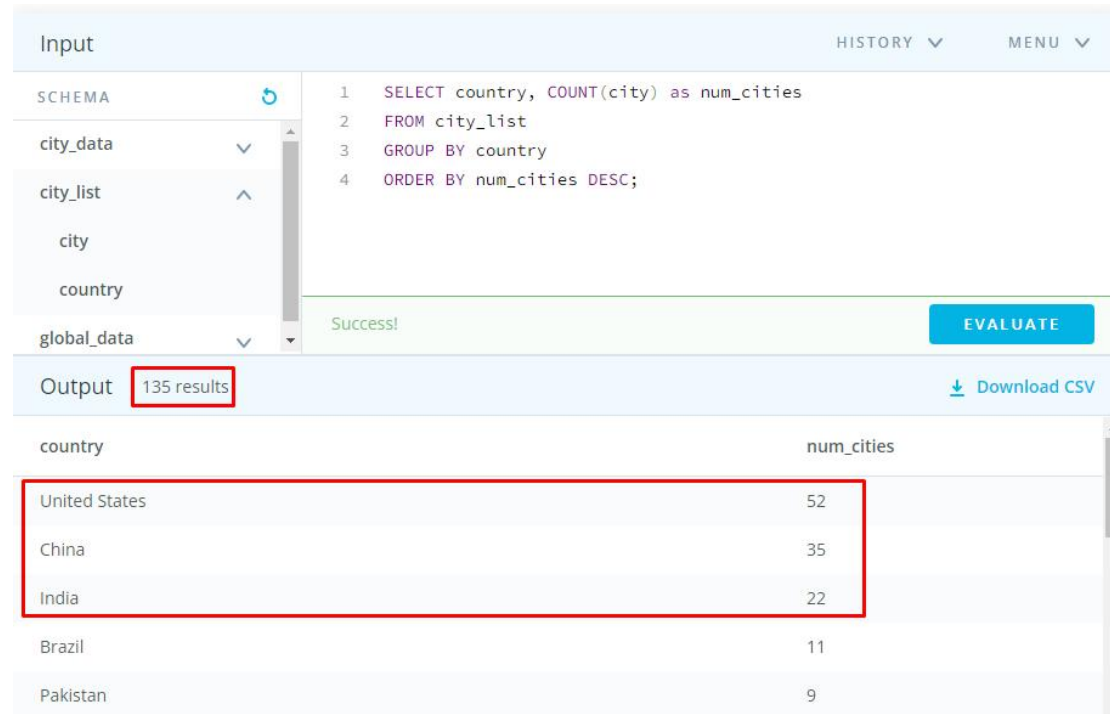
(Additional SQL queries which I am interested in)

```
SELECT country, COUNT(city) as num_cities
```

FROM city_list

GROUP BY country

ORDER BY num_cities DESC;



The screenshot shows a SQL query editor interface. On the left, there's a 'SCHEMA' panel with a tree view containing 'city_data', 'city_list', 'city', 'country', and 'global_data'. The 'city_list' table is selected. The main area shows a SQL query:

```
1 SELECT country, COUNT(city) as num_cities
2 FROM city_list
3 GROUP BY country
4 ORDER BY num_cities DESC;
```

 Below the query, there's a 'Success!' message and an 'EVALUATE' button. The 'Output' section shows '135 results' and a 'Download CSV' button. The output is a table with two columns: 'country' and 'num_cities'. The first five rows are highlighted with a red box: United States (52), China (35), India (22), Brazil (11), and Pakistan (9).

country	num_cities
United States	52
China	35
India	22
Brazil	11
Pakistan	9

Students who are living in United states, China or India and participating in this Nanodegree has higher chance to choose their exact city where they live. And from the query we can also conclude that there are 135 different countries in the table called **city_list**.

You can also check it using this query below:

SELECT DISTINCT(country)

FROM city_data;

Prepared CSV file

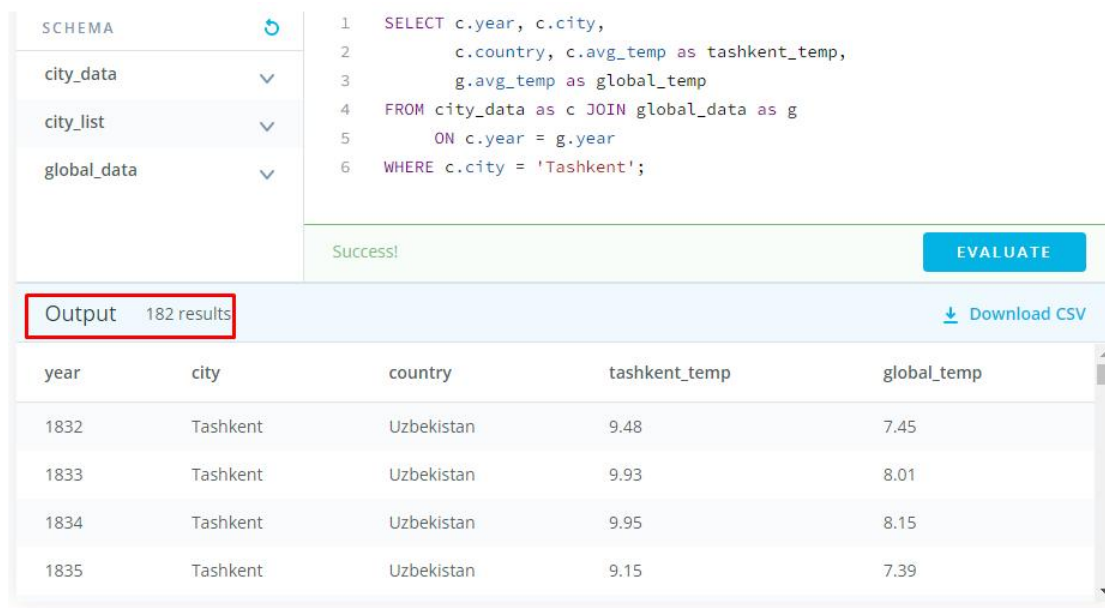
After that I prepared a CSV which is necessary for my analysis.

I used this SQL Query to get the CSV.

- (NOTE that I checked null values in my Excel sheet.

- Moreover, we can also exclude c.city and c.country as their only the Tashkent's temperature - I noticed after finishing my analysis.
- Other thing to mention is in this query I am analyzing the matching years with Tashkent's and Global's avg_temperature that's why Global years from 1750 to 1831 is not included. It starts with matching year from 1832 to 2013).

```
SELECT c.year, c.city,
       c.country, c.avg_temp as tashkent_temp,
       g.avg_temp as global_temp
FROM city_data as c JOIN global_data as g
     ON c.year = g.year
WHERE c.city = 'Tashkent';
```

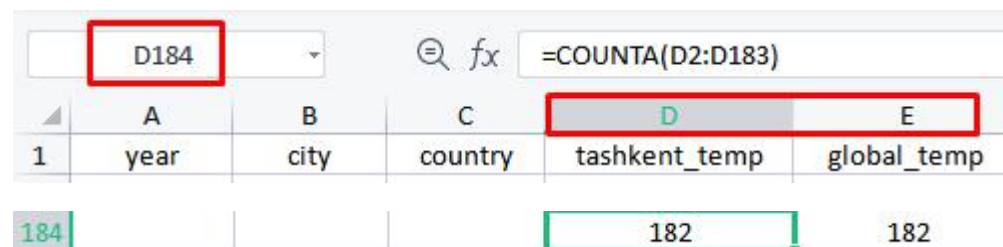


year	city	country	tashkent_temp	global_temp
1832	Tashkent	Uzbekistan	9.48	7.45
1833	Tashkent	Uzbekistan	9.93	8.01
1834	Tashkent	Uzbekistan	9.95	8.15
1835	Tashkent	Uzbekistan	9.15	7.39

I saved this CSV as tashkent_temp for my further analysis. It contains 5 attributes which are *year*, *city*, *country*, *tashkent_temp* and *global_temp*. As you can see there are 182 results which means I can explore the weather trends which changed in 182 years nearly 2 century over two trends Tashkent's and Global's temp.

Open up the CSV

After that I opened up the CSV. In order to do analysis we must also check that whether there is null(blank) values in *tashkent_temp* and *global_temp* attributes.



The screenshot shows an Excel spreadsheet with the following data:

	A	B	C	D	E
1	year	city	country	tashkent_temp	global_temp
184				182	182

The formula bar shows the formula `=COUNTA(D2:D183)` in cell D184. The cell D184 is highlighted with a red box. The cell D182 is highlighted with a green box.

You can see that we do not have any null values and so we can go to the other steps to find the moving average.

Moving Average Calculation

Moving average is statistical technique to analyze time-series data to identify patterns and trends. In order to calculate moving average we should start by calculating average value of a set of data points over certain period of time. The point of doing this thing is smoothing out short-term fluctuations and revealing longer-term trends. In this project, I calculated moving average using **Excel** tool.

I created two columns called *Tashkent_10yMA* and *Global_10yMA* (10yMA stands for 10 year moving average). I went down the 10th year(1841) and use the `AVERAGE()` function to calculate the average temperature for the first 10 years of Tashkent temperature. I also used `ROUND()` function in order to round the number for two

decimals. I also followed the same approach to find moving average for column called Global_10yMA. You can see the picture below.

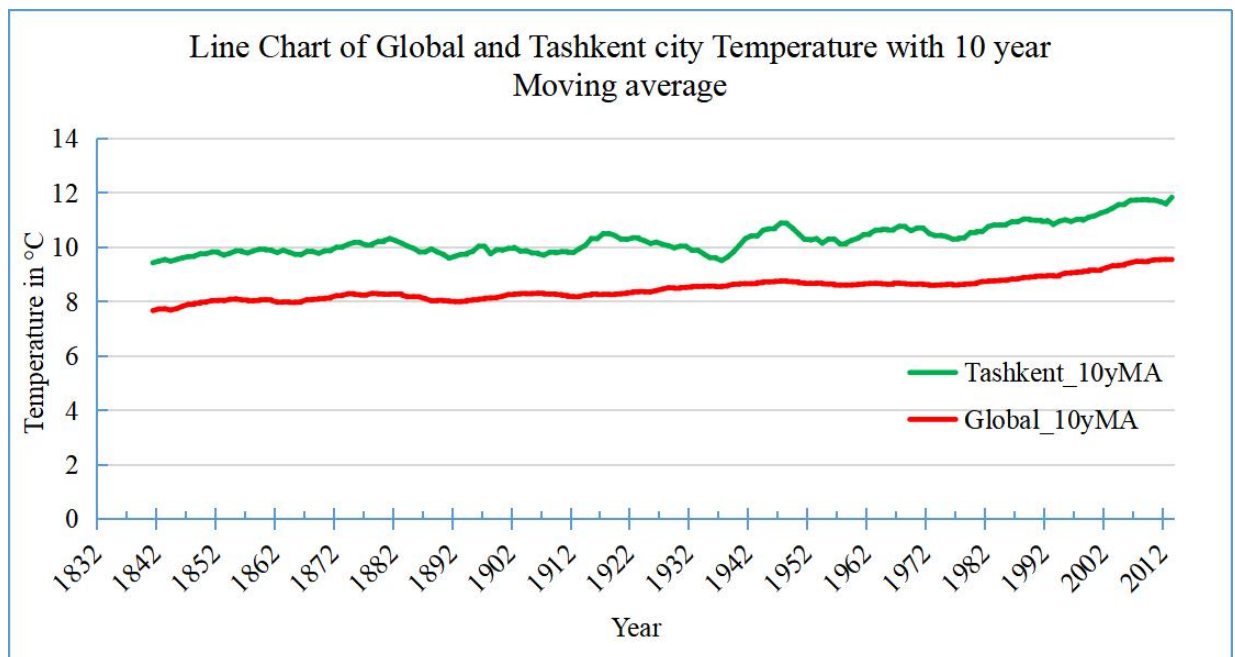
	F11		fx	=ROUND(AVERAGE(D2:D11),2)			
	A	B	C	D	E	F	G
1	year	city	country	tashkent_temp	global_temp	Tashkent_10yMA	Global_10yMA
2	1832	Tashkent	Uzbekistan	9.48	7.45		
3	1833	Tashkent	Uzbekistan	9.93	8.01		
4	1834	Tashkent	Uzbekistan	9.95	8.15		
5	1835	Tashkent	Uzbekistan	9.15	7.39		
6	1836	Tashkent	Uzbekistan	9.65	7.7		
7	1837	Tashkent	Uzbekistan	9.32	7.38		
8	1838	Tashkent	Uzbekistan	9.02	7.51		
9	1839	Tashkent	Uzbekistan	9.05	7.63		
10	1840	Tashkent	Uzbekistan	9.63	7.8		
11	1841	Tashkent	Uzbekistan	9.13	7.69	9.43	7.67
12	1842	Tashkent	Uzbekistan	10.1	8.02	9.49	7.73
13	1843	Tashkent	Uzbekistan	10.51	8.17	9.55	7.74
14	1844	Tashkent	Uzbekistan	9.38	7.65	9.49	7.69
15	1845	Tashkent	Uzbekistan	9.66	7.85	9.55	7.74
16	1846	Tashkent	Uzbekistan	10.33	8.55	9.61	7.83
17	1847	Tashkent	Uzbekistan	9.79	8.09	9.66	7.9

Key consideration to choose visualization.

After calculating moving average, I created **Line Chart** to represent the visualization of the trends of the average temperature. *The first key consideration* is the purpose of visualizing trends is to get some valuable information and also want to know the difference and similarities of two trends which are avg. Tashkent temp and avg. Global temp. I also give questions to myself like what insights I can gather from data. *Next thing* is that I identified the type of data to do my analysis. I realized that I am going to deal with time-series data. And I am using temperatures in this project. Temperature can take on any values within a range of values. It means that I am going to work with continuous data not discrete. *Next step* is that I also take into consideration the tool that I am using. It is Excel which has user-friendly interface. This gives me opportunity to make my chart informative. Moreover, I can customize my line chart to fit my

specific needs like changing fonts and colors, labels and so on. *After visualizing* my trend using Line Chart, I gave some design (adding legends) like giving colors for trends and adding axis titles, chart title for my chart to be easily seen by audience. The thing is that this chart is suitable for both technical and non-technical audience.

Line Chart



Observations

1. From the line chart we can conclude that temperature in Tashkent are greater than temperature in Global in all the time. It means my city is hotter compared to the global average temperature. There aren't any crossing points with two lines.
2. Another interesting thing is that the world is getting hotter as well as my city in an overall trend.
3. Next astonishing observation is temperature in Tashkent changed approximately from 9.5 °C to nearly 12 °C which is 2.5 °C change

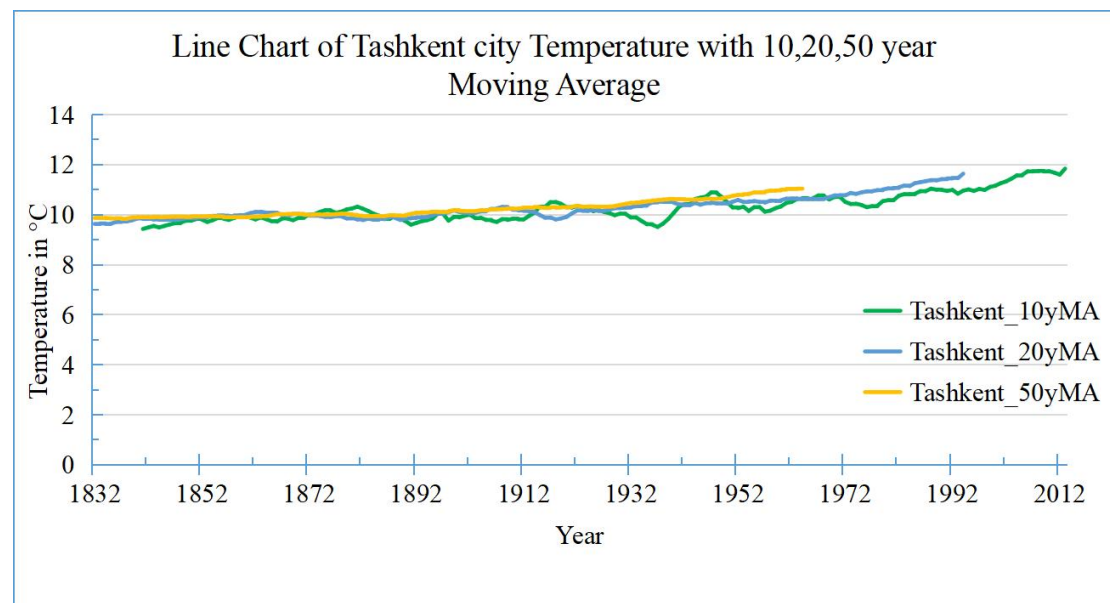
while Global temperature changed from approximately 7.5 °C to 9.5 °C which is 2.0 °C change. It means that difference between changing of temperatures for these two trends is $2.5 - 2.0 = 0.5$ °C.

4. There is also another similarity in degrees. Both of these two trends reached their maximum degrees in the last 10 year Moving average which is 2013. (see Excel file F184-186, G184-186 for more info).

5. From the 3rd observation we can analyze that the initial temperature of Tashkent city's average temperature is 9.5 °C in the first ten decades of Moving Average while the temperature of global average temperature reached 9.5 °C in its final data point.

6. Moreover, from 1912 to 1982 there are 3 fluctuations in the trend called Tashkent, while the other trend shows gradual increase in those times.

Additional visualization #1 with summary



You can see the picture above which represents the Tashkent city's average temperature in 10, 20 and 50 year Moving average. We can

say that they are exactly the same. There is some small differences in Tashkent's 10 year moving average when the year was 1937. There were some fluctuations. (I saved into the xlsx file called *tashkent_temp102050*).

Additional remarkable insight #2

This SQL query executes one beautiful city from each continent with their avg_temp corresponding to the matching years with all of these cities existed years. Unfortunately, due to the problem of the workspace, I couldn't do further analysis. But if you have successfully downloaded the table called **city_data**, you can find interesting insights of these countries moving averages and you can summarize the avg_temp of different continents. You can get valuable infos about them. After you have done with CSV file. You can open it and do same steps as I did before.

```
SELECT Asia.year, Asia.avg_temp as tashkent_temp,  
       Europe.avg_temp as paris_temp,  
       N_America.avg_temp as los_angeles_temp,  
       S_America.avg_temp as rio_de_janiero_temp,  
       Australia.avg_temp as melbourn_temp,  
       Africa.avg_temp as casablanca_temp  
FROM city_data as Asia  
JOIN city_data as Europe ON Asia.year = Europe.year  
JOIN city_data as N_America ON Asia.year = N_America.year  
JOIN city_data as S_America ON Asia.year = S_America.year  
JOIN city_data as Australia ON Asia.year = Australia.year
```

```

JOIN city_data as Africa ON Asia.year = Africa.year
WHERE (Asia.city = 'Tashkent' AND tashkent_temp IS NOT NULL)
      OR (Europe.city = 'Paris' AND paris_temp IS NOT NULL)
      OR (N_America.city = 'Los Angeles' AND los_angeles_temp IS
NOT NULL)
      OR (S_America.city = 'Rio De Janeiro' AND rio_de_janeiro_temp
IS NOT NULL)
      OR (Australia.city = 'Melbourne' AND melbourne_temp IS NOT
NULL)
      OR (Africa.city = 'Casablanca' AND casablanca_temp IS NOT
NULL);

```

You can imagine the tables attributes by looking to the picture below.

	A	B	C	D	E	F	G
1	year	tashkent_temp	paris_temp	los_angeles_temp	rio_de_janeiro_temp	melbourn_temp	casablanca_temp

I used these websites (see after the context) to choose the best beautiful city from each country. While taking best beautiful cities from websites, I also take into consideration whether these cities are existed in our database table which is called **city_data**. As everyone knows there are 6 continents not including **Antarctica**. I took the city Los Angeles from **North America**, Rio De Janeiro from **South America**, Paris from **Europe**, Tashkent (where I live) from **Asia**, Casablanca from **Africa**, Melbourne from **Australia**

<https://travelfore.com/most-beautiful-cities-from-each-continent/>

<https://theplanetd.com/most-beautiful-cities-in-europe/>

<https://www.travelessence.co.uk/australia/cities>

<https://www.attractionsofamerica.com/travel/most-beautiful-cities-usa.php>