Exploring Weather Trends- Project Submission

Bilal Karim – Udacity Data Analyst Nanodegree

Tools used for project

1) SQL for data extraction (queries and method provided below)
2) MS Excel for data visualization and analysis/profiling

Data extraction using SQL

I ran the same queries to download separate files for 2 cities. I live in Toronto, so I picked that for my first city. I picked Karachi as my comparison city because a) it is generally warmer and closer to the equator than Toronto and b) I am interested in learning more about its temperature trends. I used the following queries to extract data from the database:

SQL query used to extract Toronto data =

```
select city_data.year as Year, city_data.avg_temp as "Toronto_Avg_Temp"
from city_data
where city = 'Toronto' and country = 'Canada' order by 1
```

SQL query used to extract Karachi data =

```
select city_data.year as Year, city_data.avg_temp as "Karachi_Avg_Temp"
from city_data
where city = 'Karachi' and country = 'Pakistan' order by 1
```

SQL query used to extract global data =

```
select global_data.year as Year, global_data.avg_temp as "Global_temp"
from global_data
```

Preparing data for analysis

The first step I did was to ensure there was no duplicate data. I used the 'Remove Duplicates' functionality in Excel on the 'Year' columns in all 3 downloaded files and found no duplicate years.

I then copied the data from all 3 csv files, ensuring to match up the first years data was available for. I found 2 methods to deal with missing data on an article on Towards Data Science:

1) Remove observations entirely for which a data point is missing
2) Impute the values with either the mean, median, or the mode

Since the temperature data was missing at random, it made up a small percentage of the dataset (Karachi was missing 10% of its data and Toronto was missing 1%), and due to the risk of biasing the trend analysis with either of the measures of central tendency, I chose to use "listwise deletion" for missing values in both the Toronto and Karachi temperature sets. In the end, I was left with 197 data points to learn from. The data was arranged as follows:

Year	Global_temp	KHI_temp	YYZ_temp

I used airport codes (KHI = Karachi and YYZ = Toronto) to maintain a cleaner look across the spreadsheet.

Adding moving averages to the sheet

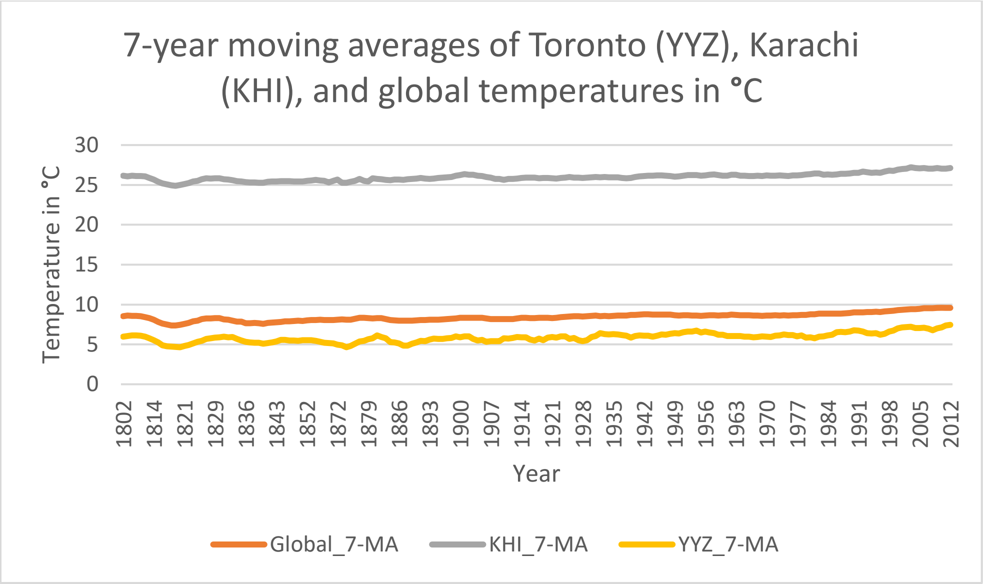
First, I added the following columns:

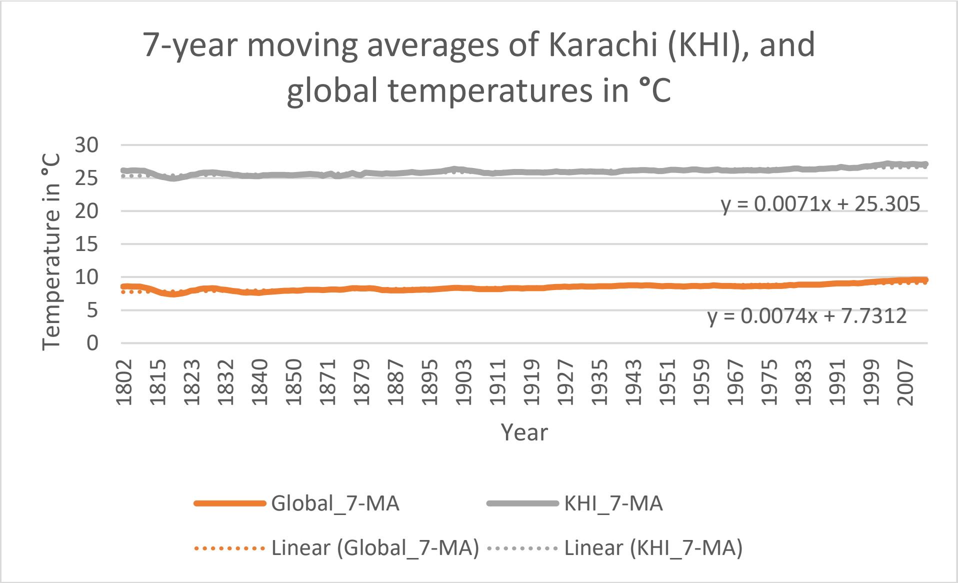
Global_7-MA	KHI_7-MA	YYZ_7-MA

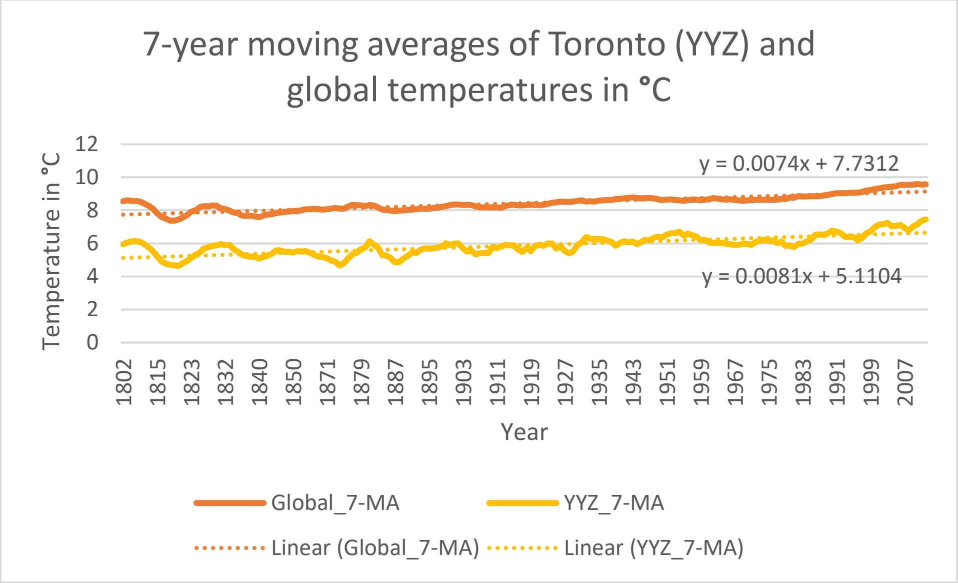
Then, I used the Excel formula 'AVERAGE' for averaging the data in 7-year increments. I rounded the moving averages to 2 decimal places by using the 'ROUND' formula. So, my 7-year moving average for global temperature from 1796-1802 was calculated as `=ROUND(AVERAGE(B2:B8),2)`, and so on. I double clicked the bottom right of each result to copy the formula for all values in that column.

To visualize the data, I plotted all the moving averages on the same chart. I used 'Select Data' in Excel to ensure that the correct data was selected as the X- and Y- values. I changed the Chart Title to '7-year moving averages of Toronto (YYZ), Karachi (KHI), and global temperatures in °C', the X-axis to 'Year', and the Y-axis to 'Temperature in °C'.

My 3 data visualizations are below:







Considerations when visualizing data

I had to keep in mind a few things when visualizing the data:

1) Since I am colorblind, I had to ensure that the colors were distinct enough to easily spot the different lines.
2) The axes and chart need to be succinct yet descriptive, so that the reader can understand the meaning of the chart in 10 seconds or less.
3) I had to make a compromise between detail and meaningfulness of the chart. I saw that when I plotted all 3 regions on the same chart, some of the details got lost because I had to zoom out the axes. Instead, I included 3 diagrams in the submission – 1 containing all 3 trend lines (Karachi, Toronto, and global) as an example, and 1 for each city compared to the global weather.

4) I also found it easier to create a narrative if I used the same colors across different charts. E.g. Using gold color for ‘YYZ_7-MA’ across 2 charts made it easier for me to refer between charts to see various detail levels.
5) Finally, adding trend lines made the charts harder to navigate and click through. In retrospect, I would have done trend line analysis on separate charts.

Observations from the data

1) On the overall, the trends show that global temperatures are rising. Using the ‘Add Chart Element’ functionality in Excel, I added a trend line for both cities and the global temperatures. The trend line for Toronto has a slope of +0.0081, the trend line for Karachi has a slope of +0.0071, and the global temperature trend line has a slope of +0.0074. This shows that moving along each of these lines, there is an average increase of these respective amounts in degrees Celsius for each year.
   The theory that global temperatures are rising overall are supported by the averages of the first and last 10 observations for each the globe, Karachi, and Toronto, as shown below:

	Global	**Karachi**	**Toronto**
Average of first 10 observations of the dataset	8.551	26.15	5.962

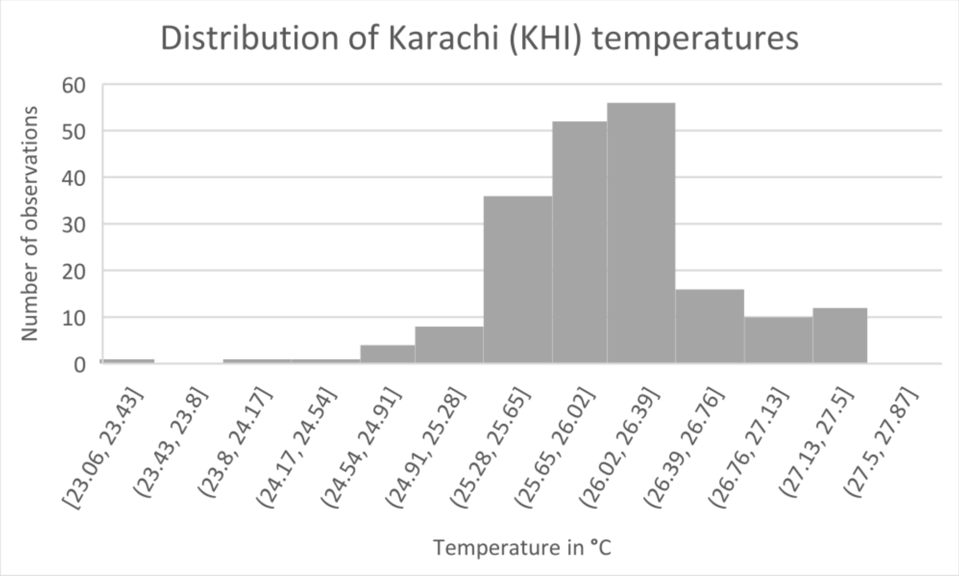
	Global	**Karachi**	**Toronto**
Average of last 10 observations of the dataset	9.556	27.082	7.359

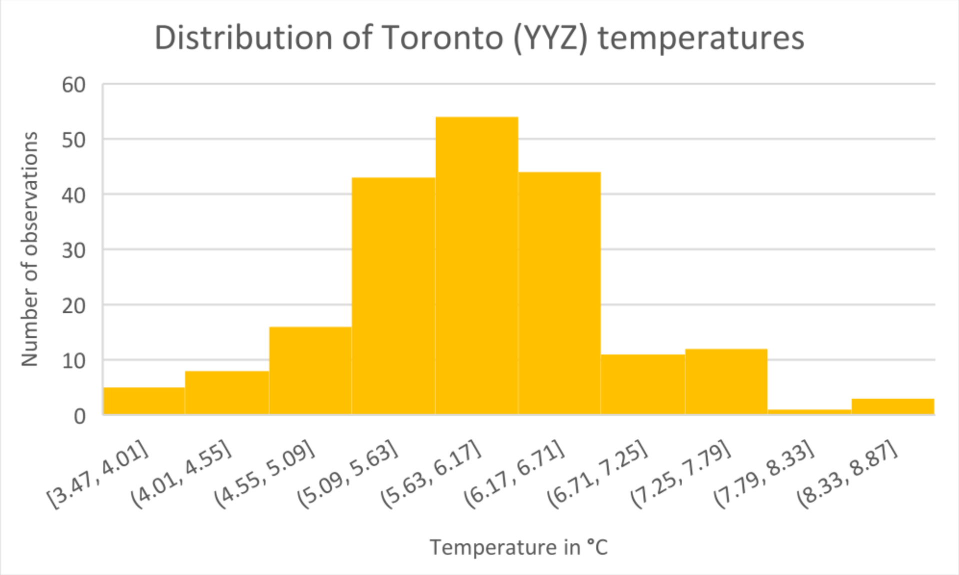
2) The average temperature in Toronto appears to be rising at a higher rate higher than the global temperature, meanwhile the temperature in Karachi seems to be rising slower than the global temperature. Despite increasing in general, the temperatures also show more cyclic tendencies by rising sharply in some years and falling in subsequent years before rising again.

3) The temperature increases in both my chosen cities are correlated with the global temperature increase. I used the “CORREL” formula to compare the correlations between ‘Toronto & Global’ and ‘Karachi & Global’ temperatures and found the following: This shows that the increase in overall global temperatures relates to the temperature increase in both cities. The closer this number is to +1, the stronger the correlation it depicts. Karachi shows a stronger correlation because the number is higher (0.78) as compared to that of Toronto (0.70).

	Correlation coefficient
Toronto/Global	0.70
Karachi/Global	0.78

4) Since both the datasets for Toronto and Karachi are normally distributed (charts are provided below), their respective standard deviations show that the temperature in Toronto varies more than the temperature in Karachi, which is comparatively more stable. I calculated the standard deviations using the formula `=STDEV.P(C2:C198)`.





5) Both cities (Toronto and Karachi) hit their minimum temperatures (3.47°C and 23.06°C respectively) between 1874 and 1875. On a similar note, both cities hit their highest temperatures in the dataset in the last 15 years (Toronto 8.66°C in 2012 and Karachi 27.5°C in 2004).

Appendix A: Full final dataset used for above analysis

Year	Global_temp	KHI_temp	YYZ_temp	Global_7-MA	KHI_7-MA	YYZ_7-MA
1796	8.27	25.89	5.48			
1797	8.51	27.03	5.21			
1798	8.67	25.59	6.03			
1799	8.51	26.14	5.62			
1800	8.48	26.05	5.85			
1801	8.59	25.52	6.23			
1802	8.58	26.45	6.46	8.52	26.1	5.84
1803	8.5	26.2	6.23	8.55	26.14	5.95
1804	8.84	26.49	6.02	8.6	26.06	6.06
1805	8.56	26.14	6.49	8.58	26.14	6.13
1806	8.43	26.06	5.7	8.57	26.13	6.14
1807	8.28	25.85	5.49	8.54	26.1	6.09
1813	7.74	25.4	5.31	8.42	26.08	5.96
1814	7.59	24.86	5.16	8.28	25.86	5.77
1815	7.24	24.92	4.52	8.1	25.67	5.53
1816	6.94	24.46	4	7.83	25.38	5.24
1817	6.98	24.69	4.12	7.6	25.18	4.9
1818	7.83	25.43	4.84	7.51	25.09	4.78
1819	7.37	24.85	5.07	7.38	24.94	4.72
1820	7.62	25.01	5.08	7.37	24.89	4.68
1821	8.09	25.62	4.88	7.44	25	4.64
1822	8.19	25.69	5.64	7.57	25.11	4.8
1823	7.72	25.4	4.86	7.69	25.24	4.93
1824	8.55	26.12	5.38	7.91	25.45	5.11
1826	8.36	25.95	6.23	7.99	25.52	5.31
1827	8.81	26.23	6.03	8.19	25.72	5.44
1828	8.17	25.73	6.82	8.27	25.82	5.69
1829	7.94	25.54	5.46	8.25	25.81	5.77
1830	8.52	25.94	6.23	8.3	25.84	5.86
1831	7.64	25.24	4.96	8.28	25.82	5.87
1832	7.45	25.42	5.9	8.13	25.72	5.95
1833	8.01	25.66	5.88	8.08	25.68	5.9
1834	8.15	25.72	6.16	7.98	25.61	5.92
1835	7.39	24.85	5.11	7.87	25.48	5.67
1836	7.7	25.3	4.27	7.84	25.45	5.5
1837	7.38	25.17	4.89	7.67	25.34	5.31
1838	7.51	25.17	4.48	7.66	25.33	5.24
1839	7.63	25.32	5.62	7.68	25.31	5.2
1840	7.8	25.44	5.88	7.65	25.28	5.2
1841	7.69	25.75	5.33	7.59	25.29	5.08
1842	8.02	25.64	5.75	7.68	25.4	5.17

1843	8.17	25.58	4.81	7.74	25.44	5.25
1844	7.65	25.24	5.8	7.78	25.45	5.38
1845	7.85	25.42	5.81	7.83	25.48	5.57
1848	7.98	25.37	5.68	7.88	25.49	5.58
1849	7.98	25.42	5.18	7.91	25.49	5.48
1850	7.9	25.44	5.45	7.94	25.44	5.5
1851	8.18	25.61	5.45	7.96	25.44	5.45
1852	8.1	25.62	5.27	7.95	25.45	5.52
1853	8.04	25.63	5.7	8	25.5	5.51
1854	8.21	25.77	5.9	8.06	25.55	5.52
1855	8.11	25.93	5.16	8.07	25.63	5.44
1856	8	24.94	4.26	8.08	25.56	5.31
1857	7.76	25.15	4.64	8.06	25.52	5.2
1871	8.12	24.1	5.26	8.05	25.31	5.17
1872	8.19	27.16	4.83	8.06	25.53	5.11
1873	8.35	26.55	4.64	8.11	25.66	4.96
1874	8.43	23.06	5.47	8.14	25.27	4.89
1875	7.86	26	3.47	8.1	25.28	4.65
1876	8.08	25.75	5.41	8.11	25.4	4.82
1877	8.54	25.96	6.48	8.22	25.51	5.08
1878	8.83	25.87	7.15	8.33	25.76	5.35
1879	8.17	25.33	5.36	8.32	25.5	5.43
1880	8.12	25.95	6.19	8.29	25.42	5.65
1881	8.27	25.93	6.31	8.27	25.83	5.77
1882	8.13	25.52	5.97	8.31	25.76	6.12
1883	7.98	25.35	4.06	8.29	25.7	5.93
1884	7.77	25.61	5.29	8.18	25.65	5.76
1885	7.92	25.61	3.83	8.05	25.61	5.29
1886	7.95	25.89	5.17	8.02	25.69	5.26
1887	7.91	25.74	5.3	7.99	25.66	5.13
1888	8.09	25.74	4.39	7.96	25.64	4.86
1889	8.32	26.02	5.96	7.99	25.71	4.86
1890	7.97	25.73	5.71	7.99	25.76	5.09
1891	8.02	25.8	6.28	8.03	25.79	5.23
1892	8.07	26.32	5.4	8.05	25.89	5.46
1893	8.06	25.3	4.83	8.06	25.81	5.41
1894	8.16	25.48	6.64	8.1	25.77	5.6
1895	8.15	25.94	5.12	8.11	25.8	5.71
1896	8.21	26.45	5.75	8.09	25.86	5.68
1897	8.29	26.09	5.89	8.14	25.91	5.7
1898	8.18	26.24	6.59	8.16	25.97	5.75
1899	8.4	26.35	5.7	8.21	25.98	5.79
1900	8.5	26.51	6.47	8.27	26.15	6.02
1901	8.54	26.18	5.7	8.32	26.25	5.89

1902	8.3	26.75	5.8	8.35	26.37	5.99
1903	8.22	25.74	5.79	8.35	26.27	5.99
1904	8.09	26.1	3.85	8.32	26.27	5.7
1905	8.23	25.55	5.21	8.33	26.17	5.5
1906	8.38	25.84	6.19	8.32	26.1	5.57
1907	7.95	25.84	4.83	8.24	26	5.34
1908	8.19	25.7	6.22	8.19	25.93	5.41
1909	8.18	25.62	5.83	8.18	25.77	5.42
1910	8.22	25.52	5.77	8.18	25.74	5.41
1911	8.18	25.5	6.43	8.19	25.65	5.78
1912	8.17	26.29	4.82	8.18	25.76	5.73
1913	8.3	25.85	6.68	8.17	25.76	5.8
1914	8.59	26.11	5.59	8.26	25.8	5.91
1915	8.59	26.24	6	8.32	25.88	5.87
1916	8.23	25.86	5.89	8.33	25.91	5.88
1917	8.02	25.59	3.91	8.3	25.92	5.62
1918	8.13	25.52	5.57	8.29	25.92	5.49
1919	8.38	25.69	6.54	8.32	25.84	5.74
1920	8.36	26.07	5.24	8.33	25.87	5.53
1921	8.57	26.2	7.75	8.33	25.88	5.84
1922	8.41	26.03	6.5	8.3	25.85	5.91
1923	8.42	25.48	5.49	8.33	25.8	5.86
1924	8.51	26.06	4.86	8.4	25.86	5.99
1925	8.53	25.95	5.63	8.45	25.93	6
1926	8.73	26.12	4.34	8.5	25.99	5.69
1927	8.52	25.63	6.1	8.53	25.92	5.81
1928	8.63	26.01	5.86	8.54	25.9	5.54
1929	8.24	25.91	5.47	8.51	25.88	5.39
1930	8.63	25.83	6.44	8.54	25.93	5.53
1931	8.72	26.29	7.56	8.57	25.96	5.91
1932	8.71	26.11	6.62	8.6	25.99	6.06
1933	8.34	25.9	6.66	8.54	25.95	6.39
1934	8.63	25.88	5.44	8.56	25.99	6.29
1935	8.52	25.69	5.6	8.54	25.94	6.26
1936	8.55	26.01	5.55	8.59	25.96	6.27
1937	8.7	25.82	6.32	8.6	25.96	6.25
1938	8.86	25.79	6.91	8.62	25.89	6.16
1939	8.76	25.75	6.14	8.62	25.83	6.09
1940	8.76	26.22	5.06	8.68	25.88	5.86
1941	8.77	26.76	6.68	8.7	26.01	6.04
1942	8.73	26.3	6.26	8.73	26.09	6.13
1943	8.76	26.17	5.19	8.76	26.12	6.08
1944	8.85	26.03	6.49	8.78	26.15	6.1
1945	8.58	25.72	5.9	8.74	26.14	5.96

1946	8.68	26.13	6.9	8.73	26.19	6.07
1947	8.8	26.38	6.18	8.74	26.21	6.23
1948	8.75	26.3	6.38	8.74	26.15	6.19
1949	8.59	26.16	7.31	8.72	26.13	6.34
1950	8.37	25.48	5.64	8.66	26.03	6.4
1951	8.63	26.29	6.26	8.63	26.07	6.37
1952	8.64	26.26	7.01	8.64	26.14	6.53
1953	8.87	26.78	7.51	8.66	26.24	6.61
1954	8.56	26.37	6.24	8.63	26.23	6.62
1955	8.63	26.19	6.97	8.61	26.22	6.71
1956	8.28	25.75	5.75	8.57	26.16	6.48
1957	8.73	25.84	6.42	8.62	26.21	6.59
1958	8.77	26.84	5.62	8.64	26.29	6.5
1959	8.73	26.33	6.4	8.65	26.3	6.42
1960	8.58	26.2	5.9	8.61	26.22	6.19
1961	8.8	25.9	6.46	8.65	26.15	6.22
1962	8.75	26.11	5.89	8.66	26.14	6.06
1963	8.86	26.63	5.49	8.75	26.26	6.03
1964	8.41	25.86	6.5	8.7	26.27	6.04
1965	8.53	26.17	5.59	8.67	26.17	6.03
1966	8.6	26.14	6.03	8.65	26.14	5.98
1967	8.7	25.94	5.66	8.66	26.11	5.95
1968	8.52	25.91	6.11	8.62	26.11	5.9
1969	8.6	26.43	6.03	8.6	26.15	5.92
1970	8.7	26.44	5.98	8.58	26.13	5.99
1971	8.6	26.21	6.37	8.61	26.18	5.97
1972	8.5	25.93	5.22	8.6	26.14	5.91
1973	8.95	26.12	7.23	8.65	26.14	6.09
1974	8.47	26.2	6.01	8.62	26.18	6.14
1975	8.74	25.87	6.75	8.65	26.17	6.23
1976	8.35	26.09	5.47	8.62	26.12	6.15
1977	8.85	26.84	6.19	8.64	26.18	6.18
1978	8.69	26.21	5.24	8.65	26.18	6.02
1979	8.73	26.28	5.9	8.68	26.23	6.11
1980	8.98	26.68	5.48	8.69	26.31	5.86
1981	9.17	26.57	6.16	8.79	26.36	5.88
1982	8.64	26.29	6.04	8.77	26.42	5.78
1983	9.03	26.1	6.78	8.87	26.42	5.97
1984	8.69	25.91	6.42	8.85	26.29	6
1985	8.66	26.34	6.13	8.84	26.31	6.13
1986	8.83	26.1	6.56	8.86	26.28	6.22
1987	8.99	26.88	7.46	8.86	26.31	6.51
1988	9.2	27.21	6.56	8.86	26.4	6.56
1989	8.92	26.25	5.7	8.9	26.4	6.52

1990	9.23	26.39	7.41	8.93	26.44	6.61
1991	9.18	26.34	7.55	9	26.5	6.77
1992	8.84	26.36	5.79	9.03	26.5	6.72
1993	8.87	27.25	5.87	9.03	26.67	6.62
1994	9.04	26.36	5.95	9.04	26.59	6.4
1995	9.35	26.63	6.38	9.06	26.51	6.38
1996	9.04	26.5	5.81	9.08	26.55	6.39
1997	9.2	26.29	6	9.07	26.53	6.19
1998	9.52	27.36	8.54	9.12	26.68	6.33
1999	9.29	27.14	7.75	9.19	26.79	6.61
2000	9.2	27.12	6.67	9.23	26.77	6.73
2001	9.41	27.28	7.76	9.29	26.9	6.99
2002	9.57	27.2	7.48	9.32	26.98	7.14
2003	9.53	26.97	6.02	9.39	27.05	7.17
2004	9.32	27.5	6.4	9.41	27.22	7.23
2005	9.7	26.8	7.22	9.43	27.14	7.04
2006	9.53	26.73	7.85	9.47	27.09	7.06
2007	9.73	27.21	7.07	9.54	27.1	7.11
2008	9.43	26.74	6.58	9.54	27.02	6.95
2009	9.51	27.41	6.28	9.54	27.05	6.77
2010	9.7	27.45	7.77	9.56	27.12	7.02
2011	9.52	27.02	7.3	9.59	27.05	7.15
2012	9.51	26.75	8.66	9.56	27.04	7.36
2013	9.61	27.21	8.46	9.57	27.11	7.45