Visual Data Science
Assignment 2

Fabian Pechstein
00726104

January 2021

1 Introduction

1.1 Dataset

I selected the Statewide Integrated Traffic Records System (SWITRS) [1] to be used for this assignment, that is available trough Kaggle[1] or Zenodo [2].

The data set comes from California Highway Patrol and describes traffic collisions, involved parties and the surrounding conditions from 2001 to 2020. Three data tables are available in a SQLite database:

- *collisions*: information about collision (geo-location, timestamp, vehicle type etc.) - 9.172.565 records
- *parties*: information about involved people - 18.178.069 records
- *victims*: information about injuries - 9.463.554 records

The database file is around 6 GB in total, for convenience records have been exported to CSV-files year-wise, which will be included in the dashboard prototype.

1.2 Exploration

1.2.1 Cause Distribution of Bicycle Collisions

Due to the large number samples, I tried to narrow down my investigations by selecting only accidents that involved bicycles and look into the causes of that accident. We can see that that GPS data seems to be not that properly maintained (see figures 1, 2 and 3), or GPS devices have not been widely available before 2007. One detail that also becomes apparent, when we look at at the right hand side of all the figures, is the drop of most collisions types due to spread of SARS-COVID 19 and lockdown starting end of 2020. The top five causes respectively in comparison in each set are visible in table 1.

[1] `https://www.doi.org/10.34740/kaggle/dsv/1671261`
[2] `https://zenodo.org/record/4284843`

no.	with spatial info	without spatial info
1	wrong side of road	wrong side of road
2	improper turning	automobile right of way
3	speeding	improper turning
4	automobile right of way	traffic signals and signs
5	dui	unknown

Table 1: Top five collision causes in samples with and without spatial information.

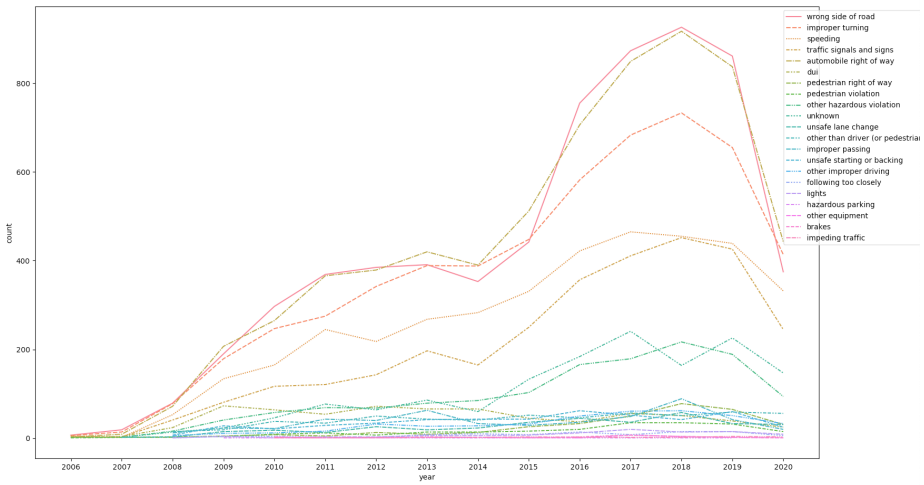

Figure 1: Bike collisions split by accidents causes with spatial information set

2 Analysis

2.1 Does the severity of the bicycle accident influence the lack of spatial information?

The data-set size of the non-spatial info group is considerably larger ($n = 220248$) in comparison to the sets without long/lat data ($n = 31989$). As mentioned before, it might we likely that the availability of GPS chips etc might explain missing data features before 2007, however if take a look at figures 5 and 6, which show the total counts of injured or killed victims split by availability of long/lat data respectively, Its visible that in the case of the occurrence of killed victims (after 2017) we see that these records bear position data (figure 6).

2.1.1 Injured cases with or without position data

Hypothesis H_0 **1** *The distributions of all samples are equal*

Hypothesis H_1 **1** *The distributions of one or more samples are not equal*

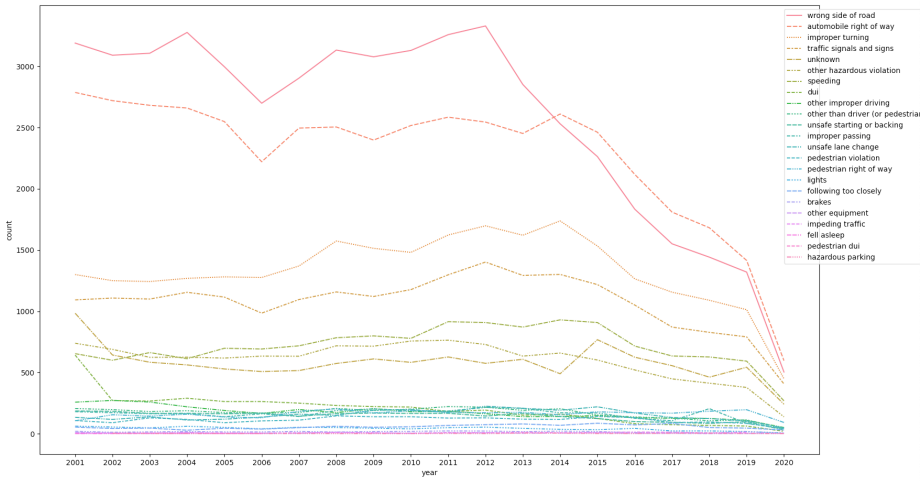

Figure 2: Bike collisions split by accidents causes without spatial information set

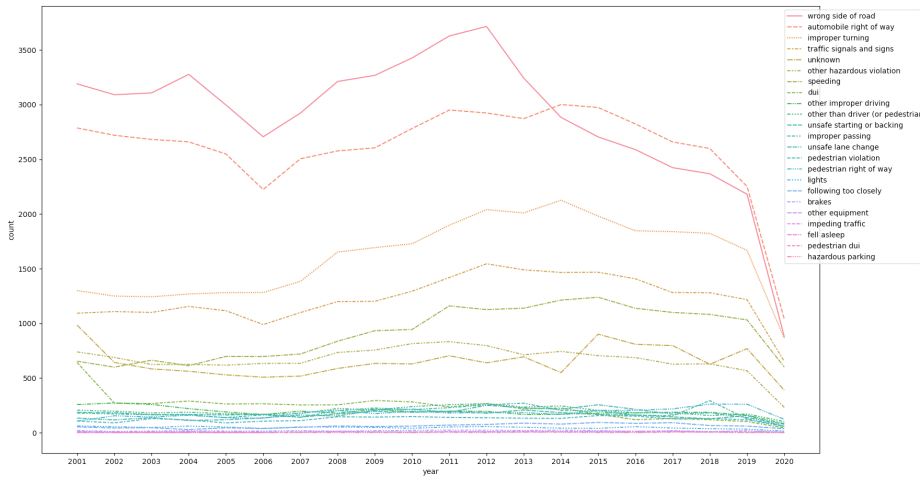

Figure 3: Bike collisions total counts split by accidents causes.

As visible in figure 4, the distribution is most likely not normally distributed, thus a Kruskal-Wallis H Test was used to test for equal distribution between the two data-sets. With $1.317, 0.18792$ we dont have enough evidence to reject H_0.

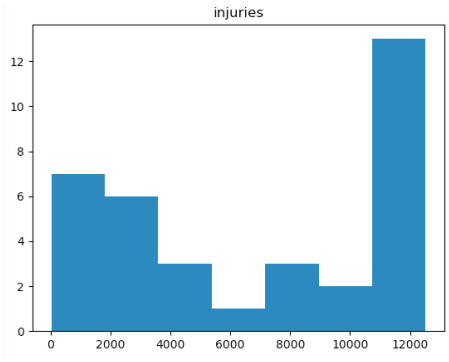

Figure 4: Histogram of yearly injury counts

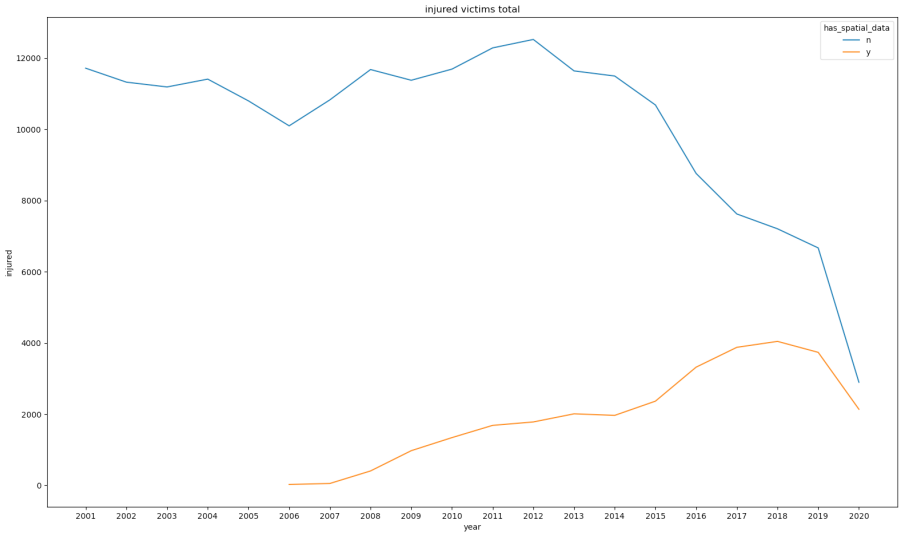

Figure 5: Comparison between the total numbers of injured victims grouped by having or lacking position information, Not significant ($p = 0.18792$)

2.1.2 Killed cases with or without position data

Hypothesis H_0 **2** *The distributions of all samples are equal*

Hypothesis H_1 **2** *The distributions of one or more samples are not equal*

Again Kruskal-Wallis H Test was used to test for equal distribution between the two data-sets. With $366.242, 0.00000$ we have **significant** evidence to reject H_0.

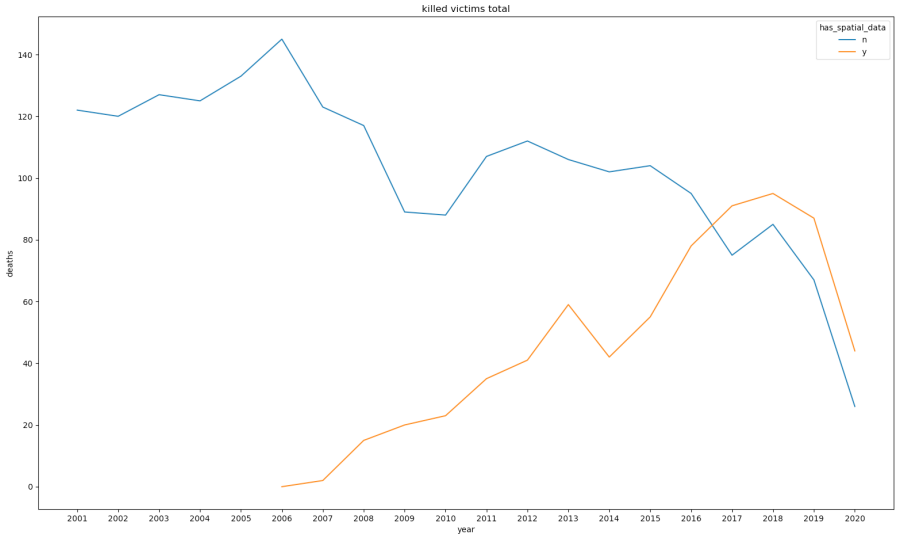

Figure 6: Comparison between the total numbers of killed victims grouped by having or lacking position information & statistically significant ($p = 0.00000$)

2.2 when do bicycle accidents occur? is there a seasonal difference?

When we look at the distributions of timestamps of bike collisions (see figure 7). The oddity of no collision occurrences between 12:00 and 13:00 cant be explained for now, but might be due to erroneous data, or by pure chance. Otherwise we see two distinctive peaks in the morning (7:00 to 9:00) and in the afternoon (15:00 to 18:00)

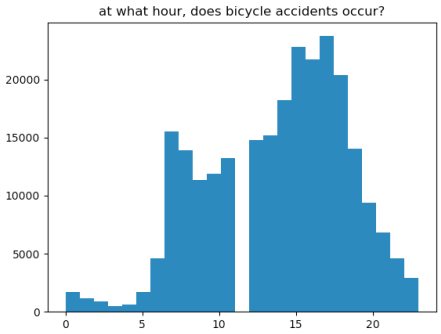

Figure 7: collision hours histogram

Well, if we take a look at the average collision times per month (see figure 8), the morning hours appear to be more dangerous in September and October,

whereas the afternoon is throughout busy claiming victims, except for December.

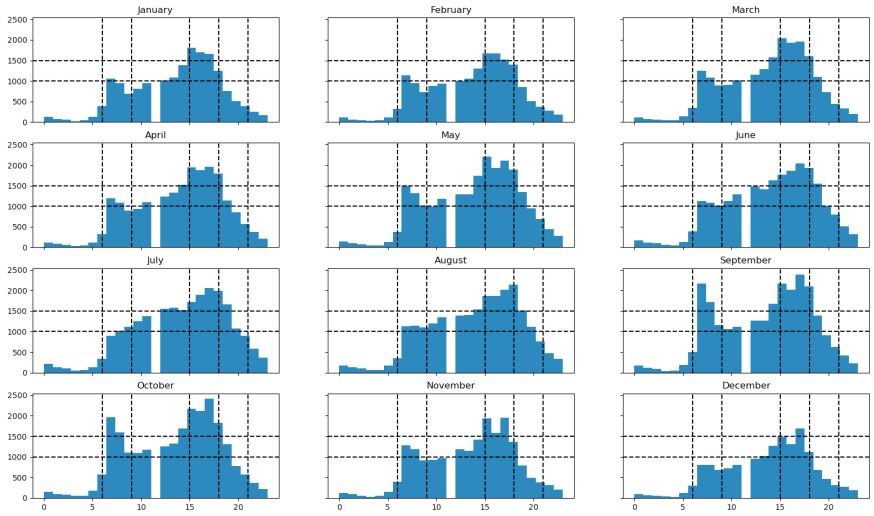

Figure 8: collision hours by month.

Again to test for statistical significance, lets assume for each of possible Month-pairs the following hypothesis:

Hypothesis H_0 **3** *the means of the samples are equal.*

Hypothesis H_1 **3** *the means of the samples are unequal.*

A student T Test has been used to determine if months are significantly different.Table A.1 for the respective p-value. A different representation can be seen in figure 9, significant values are black. By rejecting H_0 with $p > 0.05$ we found enough evidence, that both samples come from different distributions.

2.3 Is the cause of the accident related to the time?

In the same way it could become interesting to get an insight between the hour and the cause of the accident. if we group the data accordingly and take a look at the histogram of crash times, we can see that the proportion of the top six causes are somehow different in counts as well as in the observed time they occurred (see figure 10).

To test for statistical significance, lets assume for each of possible Cause-pairs the following hypothesis:

Hypothesis H_0 **4** *the distributions of both samples are equal*

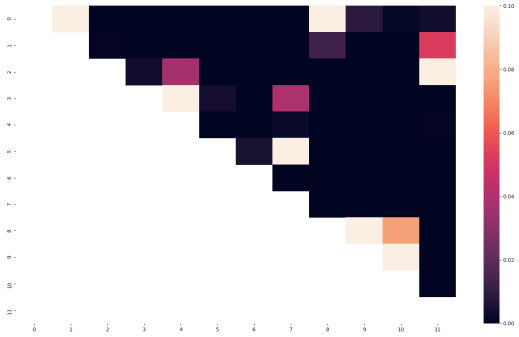

Figure 9

1	wrong side of road
2	automobile right of way
3	improper turning
4	speeding
5	traffic signals and signs
6	dui

Table 2: Top 6 accident categories

Hypothesis H_1 **4** *the distributions of both samples are not equal.*

Which gets verified using a Mann-Whitney U Test, since we assume a normal distribution from the looks of the histograms in figure 10.

From the test statistic results (see. tables A.4 and A.3) it becomes apparent the distributions are in fact quite different, except 'wrong side of road' and 'improper turning' where $p = 0.21087$ supports H_0.

3 Interpretation

Due to the smaller proportion of collisions involving bicycles in comparison to number of all collisions, we can fairly assume that Californians are not that fond of bike riding. When we look at the time, when collisions occur we can see that the two peaks we see in figure 7 correlates with commuting hours, thus is probably related to the overall traffic volume for which we have no data. Looking at the seasonal differences (figure 8) I can't really explain the observed shifts between months, but I assume it involves the weather or climate as well as holiday or leisure capabilities.

Lastly, when comparing the times grouped by the top six causes, we can observe that the distribution for 'wrong side of road' and 'automobile right

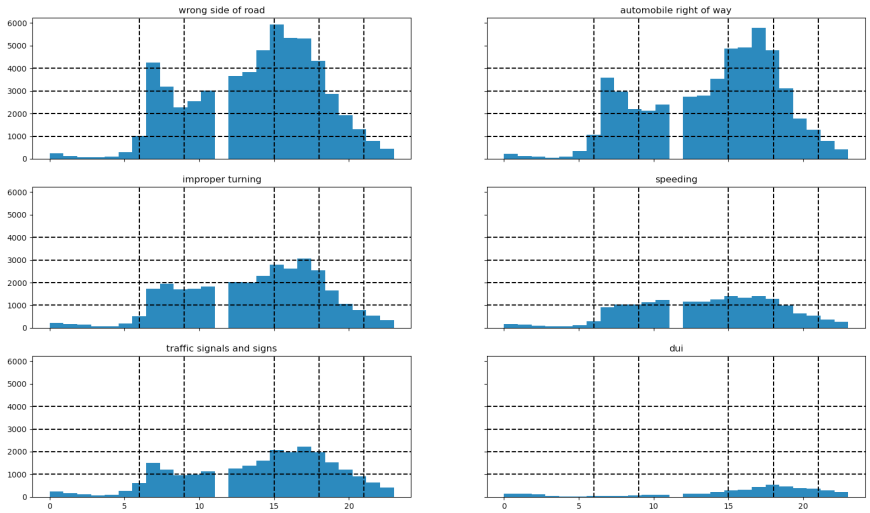

Figure 10

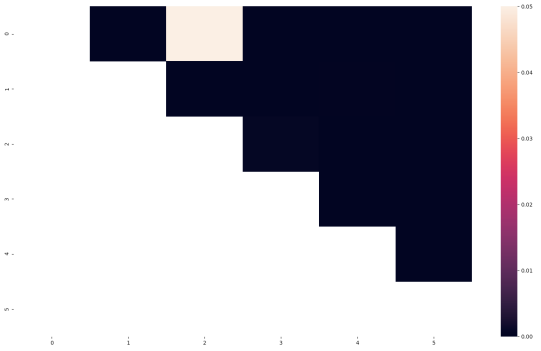

Figure 11

of way' fairly matches the average timing histogram, which makes sense when looking of the overall proportion. However, since the commuter peaks are also clearly visible, again these causes could be again related to the overall traffic volume during that time.

4 Dashboard

4.1 Personas

Probable users are bicyclists interested in historic collisions in their neighbourhood or on their commuting paths or common pitfalls.

Persona 1 Greg, 43 years old, lives in the suburbs. In his leisure time he is an avid road cyclists, trying to amass miles together with his cycling club. With the increase of traffic of the last few years he is more and more worried about getting involved in a fatal collision. When trying out new routes , he wants to check against historic hot spots for bicycle collisions on his route.

Persona 2 Anna, 25 years old, living and working in downtown L.A.. She tries to use the bike as often as possible, as she tries to reduce her carbon footprint as suggested by her yoga friend she has a crush on. After experiencing a collision with a car, she tries to raise public awareness and to support her claims about the dangers in traffic.

Persona 3 Paul (38) and Sheryl(37), parents of a 12 and 14 year old. Their kid's way to school is rather short and the kid's favourite way of travel is by bike, which fosters independence in children (or so the parents have read), however a dangerous crossing keeps them worried. They want to gather evidence to push town hall to make the school way safer.

4.2 Interactions

Possible interactions with the data involve and would be required by my personas:

- Select a data and time range to filter collisions
- Focus a area of interest on a map
- Filter by collision cause
- Filter by fatalities
- get an overview of collision causes and their distribution in the area of interest.

Based on the requirements described in 4.1 the dashboard would be most probably of the *Communication* type, with a whiff of *Decision Making*.

4.3 Rash Board

the implementation is based on Python using dash and plotly.

the inputs placed at the upper part can be used to pre-select data.

- range slider for the year
- range slider for the hour
- involved crash categories
- including/excluding fatal collisions

Data points that fulfil the defined criteria are used to fill the various plots (see 12)

- map: data points pop up at the location of the collision. colour represent the date time, circle size the severity of the crash
- pie charts show the distribution of involved crash causes for overall collisions, only collisions with injuries and fatal collisions
- the histogram shows the average distribution of collisions for the day time
- the scatter plot show the number of weekly collisions over the selected date range

4.3.1 Interactive brushing and linking

Is at the current version unfortunately not properly working yet, and remains to be seen if that's fixable until the presentation. Most important feature would be the selection of an area on the map to update the remainign plots.

References

[1] California traffic collision data from switrs. `https://www.doi.org/10.34740/kaggle/dsv/1671261`. Accessed $13^{th} January$.

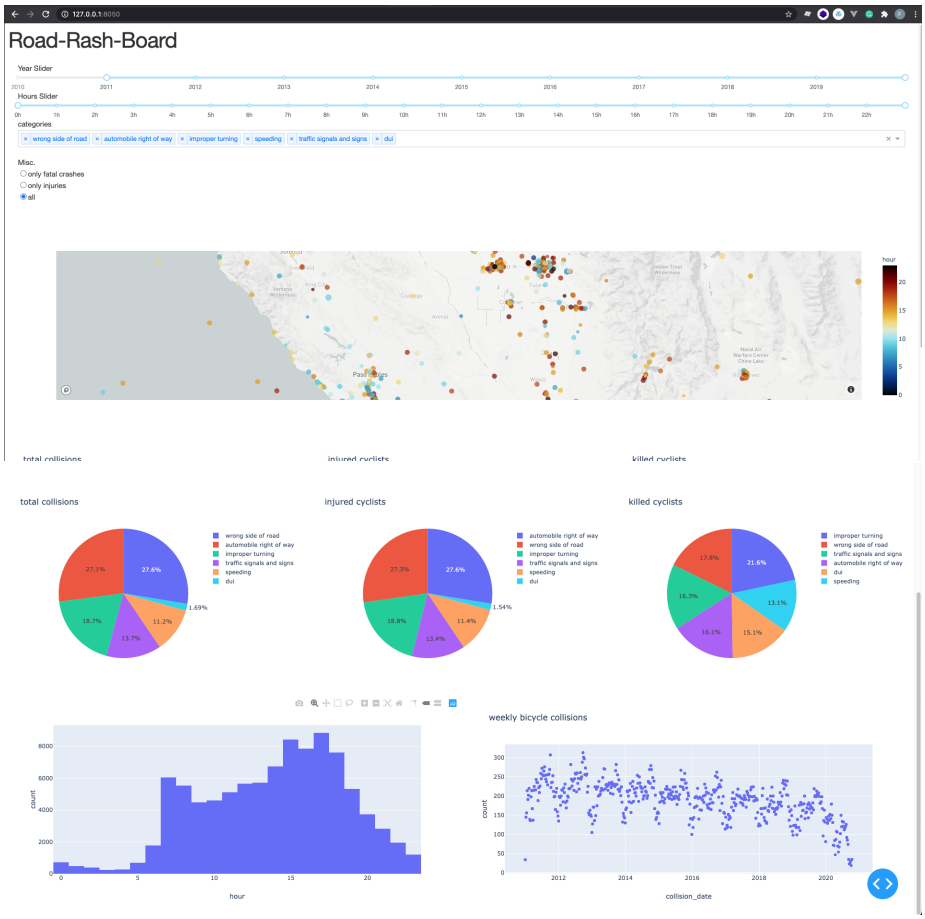

Figure 12: Rash Board

A Appendix

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Jan		0.32057	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.14929	0.00865	0.00224	0.00361
Feb			0.00043	0.00000	0.00000	0.00000	0.00000	0.00000	0.01279	0.00024	0.00005	0.05269
Mar				0.00355	0.03698	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.15813
Apr					0.38357	0.00421	0.00000	0.03895	0.00000	0.00000	0.00000	0.00004
May						0.00016	0.00000	0.00296	0.00000	0.00000	0.00000	0.00077
Jun							0.00512	0.41935	0.00000	0.00000	0.00000	0.00000
Jul								0.00030	0.00000	0.00000	0.00000	0.00000
Aug									0.00000	0.00000	0.00000	0.00000
Sep										0.20790	0.07626	0.00001
Oct											0.56622	0.00000
Nov												0.00000

Table A.1: p-values from T-Test

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Jan		-0.993	-4.570	-7.370	-6.630	-10.149	-12.836	-9.397	1.442	2.626	3.056	-2.911
Feb			-3.523	-6.314	-5.562	-9.078	-11.754	-8.327	2.490	3.672	4.059	-1.937
Mar				-2.916	-2.086	-5.792	-8.588	-5.001	6.373	7.626	7.846	1.411
Apr					0.871	-2.862	-5.653	-2.065	9.358	10.622	10.708	4.123
May						-3.779	-6.612	-2.972	8.604	9.884	9.998	3.364
Jun							-2.799	0.808	12.336	13.618	13.565	6.798
Jul								3.614	15.211	16.506	16.319	9.393
Aug									11.539	12.821	12.803	6.064
Sep										1.259	1.773	-4.470
Oct											0.574	-5.625
Nov												-5.930

Table A.2: stat-values from T-Test

	1	2	3	4	5	6
1		p=0.00000	p=0.21087	p=0.00000	p=0.00000	p=0.00000
2			p=0.00000	p=0.00000	p=0.00035	p=0.00000
3				p=0.00074	p=0.00000	p=0.00000
4					p=0.00000	p=0.00000
5						p=0.00000

Table A.3: stat-values from Mann-Whitney U Test

	1	2	3	4	5	6
1		stat=1442999967.500	stat=922944656.500	stat=512411826.000	stat=672024924.500	stat=92544615.500
2			stat=801559922.500	stat=446458087.500	stat=630081216.500	stat=87759952.500
3				stat=286091288.500	stat=372880001.000	stat=51859066.500
4					stat=207937895.500	stat=29298364.000
5						stat=43263952.000

Table A.4: stat-values from Mann-Whitney U Test