HHU@ BTW 2017 Data Science Challenge SDSC17

Alexander Askinadze & Matthias Liebeck

Heinrich-Heine-Universität Düsseldorf
{askinadze, liebeck}@cs.uni-duesseldorf.de

HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

Outline

1. Introduction
2. Analysis & Statistics
3. Visualization
4. „How to travel safely in NYC?“
5. Conclusion & Future Work

Introduction

Outline

- **Objective of the challenge**: Given a data set about car accidents in New York City, the participants were asked to explore and analyze the data set. Some questions and tasks were suggested, like for instance:
  1. Where are dangerous spots?
  2. Where are accident-free spots?
  3. Visualize the data
  4. Create an animation of the development over time
  5. Descriptive statistics and correlations, such as:
     - What types of accidents occur?
     - Are there connections between accidents and large public events?
     - What factors influence accidents?
- **Main question:** How large is the potential for avoiding accidents?

Dataset

- The primary dataset is called **NYPD Motor Vehicle Collisions**
- Size: 988k rows, each describing a reported vehicle collision
- 29 columns:
  - time and date of the accident
  - geocoordinates of the nearest intersection (and also street names)
  - number of injured *pedestrians*, *cyclists*, and *motorists* (=> summed as number of persons injured)
  - number of killed *pedestrians*, *cyclists*, and *motorists* (=> summed as number of persons killed)
  - contributing factor (e.g., following too closely or brakes defective) of the involved vehicles (up to 5 factors; 1 per vehicle)
  - vehicle type (e.g., passenger vehicle or SUV) of the involved vehicles (up to 5 vehicles)

Cloud Technology

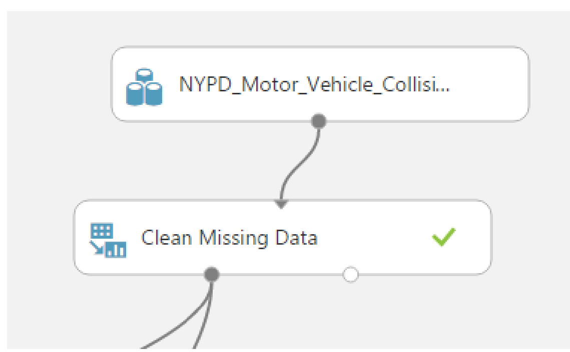
- The participation in the challenge requires the usage of cloud technologies.
- We decided to use Microsoft‘s cloud technology called Azure.

- To make our presentation a little bit more interactive, we decided to make most of the cooler stuff directly available:
  - **btw2017-dsc-hhu.azurewebsites.net/**
  - or via an URL shortener: **bit.do/hhu-btw2017**

Analysis & Statistics

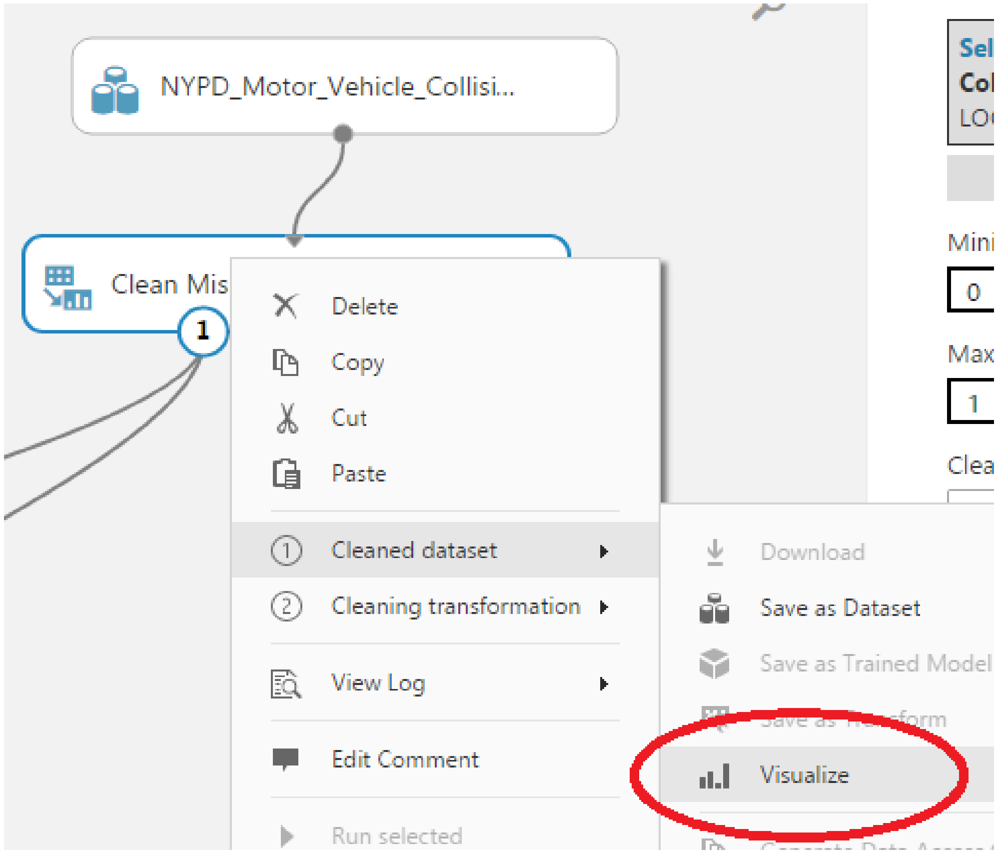
Microsoft Azure Machine Learning Studio / Azure ML

- Microsoft currently offers a free cloud based machine learning platform called **Azure ML**.
- Since Azure ML provides drag & drop functionality to process the dataset and perform machine learning operations, we started by uploading the dataset into Azure ML.
- Afterwards, we filtered the dataset from 988k rows down to 770k rows that include geocoordinates.



Azure's Visualize Function

- Azure ML contains some ready to use analysis functions:



bit.do/hhu-btw2017

Azure‘s Visualize Function

- Azure ML contains some ready to use analysis functions:

rows 769878 columns 29

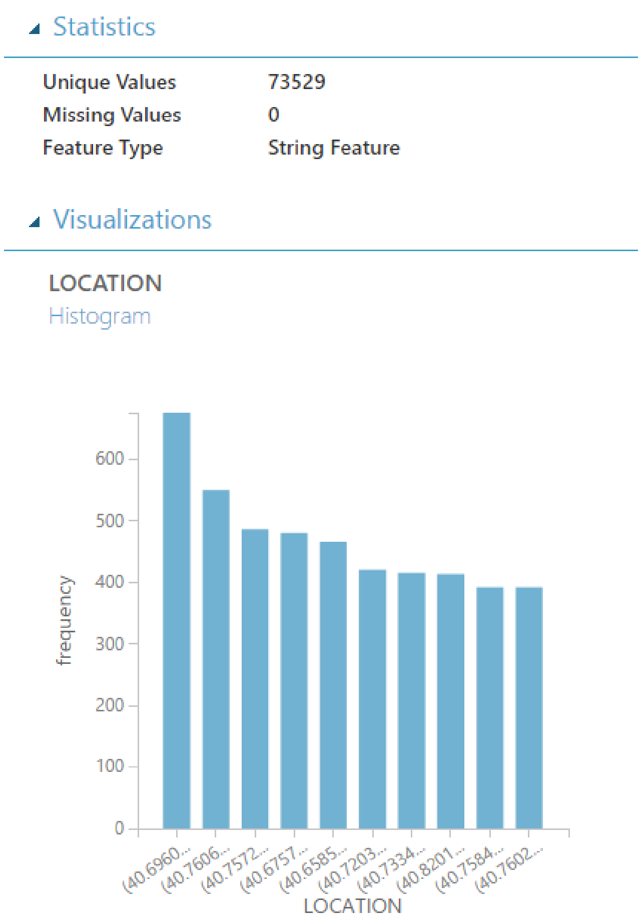
view as

DATE	TIME	BOROUGH	ZIP CODE	LATITUDE	LONGITUDE	LOCATION	ON STREET NAME	CROSS STREET NAME	OFF STREET NAME	NUMBEF OF PERSON INJURED
09/25/2016	17:09			40.841188	-73.932142	(40.8411882, -73.9321417)				0
07/18/2016	18:30	STATEN ISLAND	10312	40.549883	-74.167417	(40.5498825, -74.1674169)	RICHMOND AVENUE	LAREDO AVENUE		0
07/18/2016	18:45	QUEENS	11420	40.666096	-73.817412	(40.6660962, -73.8174115)	150 AVENUE	123 STREET		0
07/18/2016	18:48	MANHATTAN	10022	40.762913	-73.9698	(40.7629131, -73.9697999)	EAST 59 STREET	PARK AVENUE		0
07/18/2016	18:50	QUEENS	11434	40.686795	-73.7812	(40.6867946, -73.7812)			166-06 116 AVENUE	0
07/18/2016	18:50	BROOKLYN	11210	40.63406	-73.946807	(40.63406, -73.946807)	GLENWOOD ROAD	EAST 31 STREET		1
08/25/2015	19:00			40.732941	-73.920382	(40.7329414, -73.9203819)				0
11/10/2016	16:11	BRONX	10458	40.859874	-73.893216	(40.8598744, -73.8932165)	WEBSTER AVENUE	EAST 188 STREET		0
11/10/2016	16:11	BRONX	10467	40.878745	-73.872545	(40.8787451, -73.8725452)	EAST GUN HILL ROAD	DECATUR AVENUE		0
11/10/2016	16:11	BROOKLYN	11208	40.662514	-73.872007	(40.6625139, -73.8720068)	WORTMAN AVENUE	MONTAUK AVENUE		0
11/10/2016	17:11	BROOKLYN	11215	40.666845	-73.99486	(40.6668449, -73.9948598)	3 AVENUE	PROSPECT AVENUE		0

Azure's Visualize Function

Attribute **location**:

- We instantly see that:
  - there are ~73.5k unique dangerous spots
  - that approximately 700 accidents happened at the most dangerous spot



bit.do/hhu-btw2017

Azure's Visualize Function

Attribute **persons injured**:

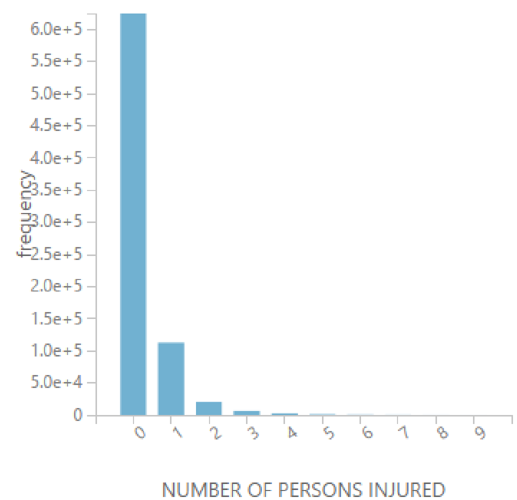
- We instantly see that:
  - no persons were injured in ~650k (84.4%) of the reported accidents
  - ~ 13% of the accidents resulted in the injury of one person

Statistics

Unique Values	25
Missing Values	1
Feature Type	String Feature

Visualizations

NUMBER OF PERSONS INJURED
Histogram



Azure's Visualize Function

Attribute **persons killed**:

- We see that:
  - Fortunately, deaths of persons are rare!
  - The maximum amount of car accident related deaths is 5

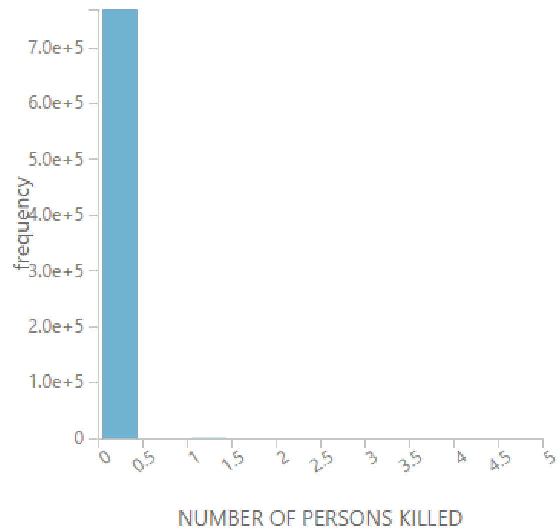
Statistics

Mean	0.0012
Median	0
Min	0
Max	5
Standard Deviation	0.0361
Unique Values	6
Missing Values	2
Feature Type	Numeric Feature

Visualizations

NUMBER OF PERSONS KILLED

Histogram



bit.do/hhu-btw2017

Azure's Visualize Function

Attribute **number of cyclist injured**:

- We have outliers!

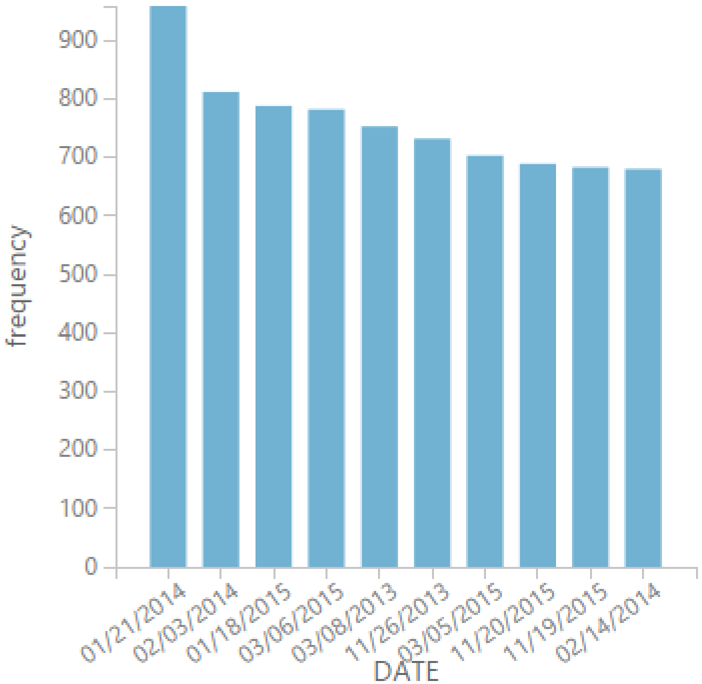
Statistics

Mean	4.2959
Median	0
Min	0
Max	3291249
Standard Deviation	3751.0259
Unique Values	8
Missing Values	1
Feature Type	Numeric Feature

Azure's Visualize Function

Attribute **date**:

- We group the accidents by date, sort the number of accidents in descending order and look for possible correlations:
  - 01/21/2014: blizzard in NYC[1]
  - 02/03/2014: day after Super Bowl 2014



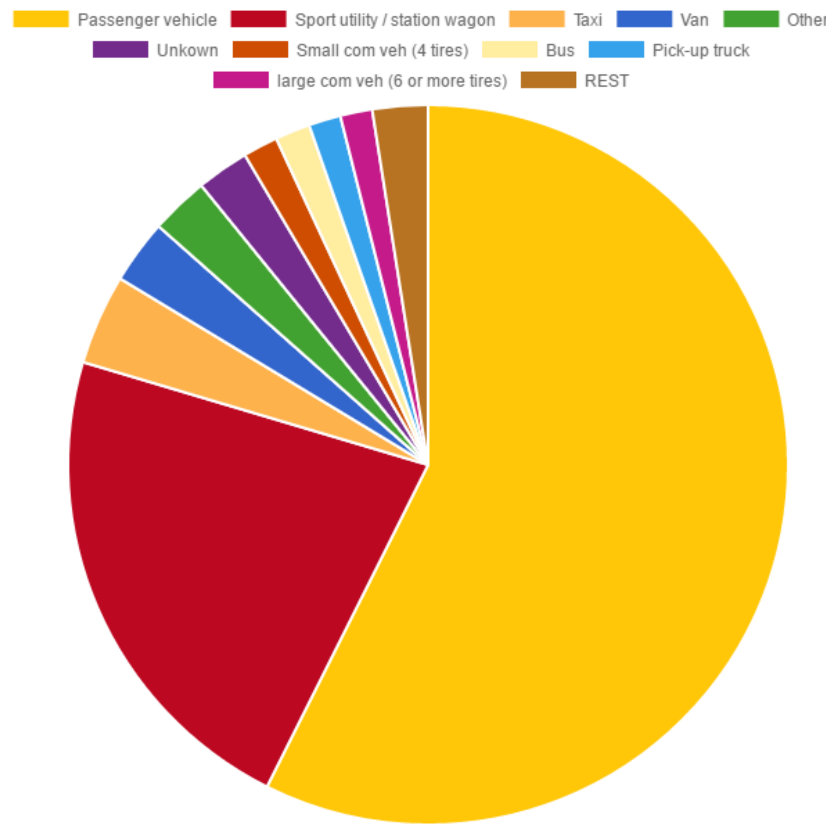
1 http://www.stuttgarter-zeitung.de/inhalt.schneechaos-in-den-usa-blizzard-legt-new-york-und-washington-lahm.6c859b7c-819a-4492-90cd-26c7e22b05bc.html

Other Visualizations

Attribute **vehicle type code 1**:

- Passenger vehicles account for most of the accidents.
- In about 25% of the accidents, SUVs or station wagons are involved.
- 0.5% involve motorcycles (not shown in the chart)
- 0.015% involve bicycles (not shown in the chart)

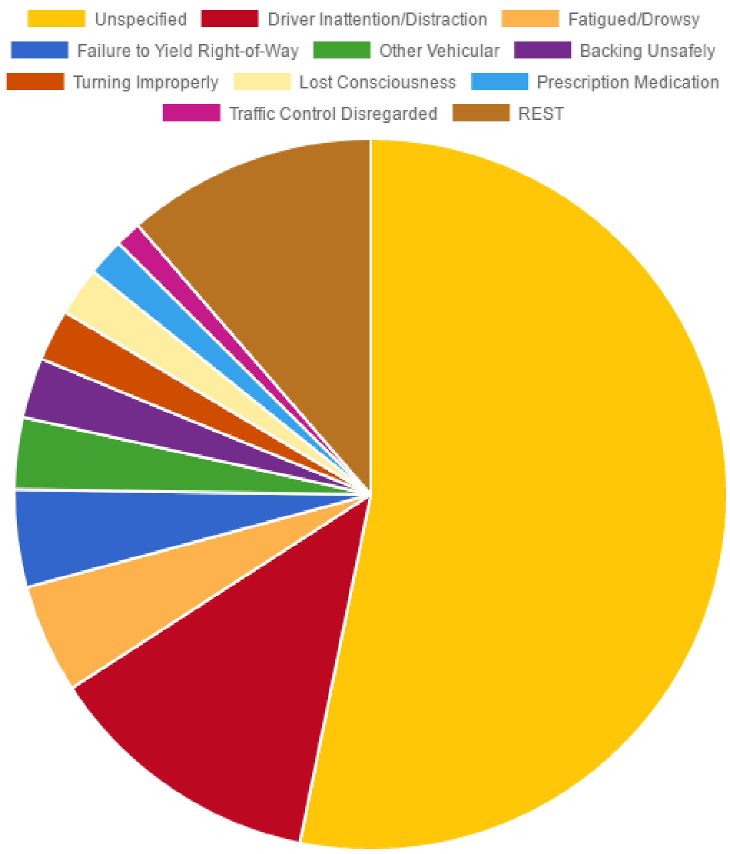
=> Don‘t drive a car, use a motorcycle or a bicycle instead ;-)



bit.do/hhu-btw2017

Attribute **contributing factor vehicle 1**:

- Regarding the question "What factors influence accidents?":
  - Over 50% of the accidents do not have a contributing factor in their accident report.
  - The most common (specified) cause is *distraction*, followed by being *fatigued*.

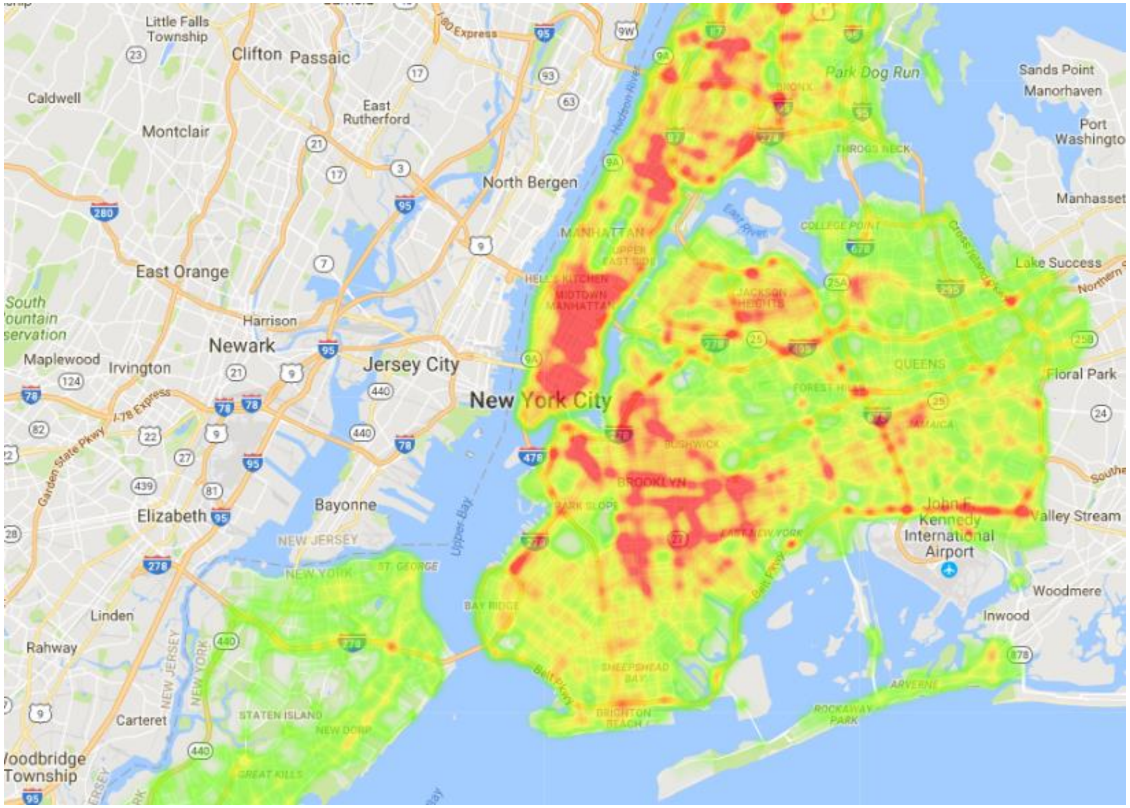


bit.do/hhu-btw2017

Visualization

Heatmaps

- Let‘s move on to some cooler analysis in the form of visualizing geocoordinates on a map.
- We plotted the accidents on a heatmap with Google Maps. The colour of a spot indicates the amount of accidents.



Heatmaps

- The means of transportation (on foot, bycycle, car, or all combined) and the timeframe can be selected.
- Let‘s switch to http://btw2017-dsc-hhu.azurewebsites.net and try it out.

Heatmaps - Animations

- We also include a filter to select a timeframe between a start and an end date.
- In order to better understand how the amounts of accidents in a certain area develop over time, we created animations for our heatmap:
  - Visualization per day
  - Visualization per day and hour
  - Visualization per hour
- We prepared a video that shows how the amount of accidents for pedestrians develops during the day.

„How to travel safely in NYC?“

„How to travel safely in NYC?"

- Up until now, we analyzed the dataset and visualized it.
- In order to contribute to the "potential for avoiding accidents", we developed a navigation software that utilizes the NYPD Motor Vehicle Collisions dataset to detect dangerous areas and suggest routes through NYC that go around these areas.
- Scenario: Given a start location and an end location, we want to plot a route that is as safe as possible while respecting the means of transportation $t \in \{pedestrian, bicycle, car\}$
- 3 external APIs:
  - Routing: HERE (navigation software, owned by Audi, BMW, and Daimler)
  - Geoencoding: Google API
  - Visualization: Google Maps

„How to travel safely in NYC?“

Workflow for the **fastest** route:

1. Enter a start and an end location
2. Choose the means of transportation
3. Geoencode the entered adresses with Google API (works better than the geoencoding from HERE)
4. Get the route via the HERE API, JSON response of GPS coordinates
5. Plot the returned GPS coordinates on Google Maps

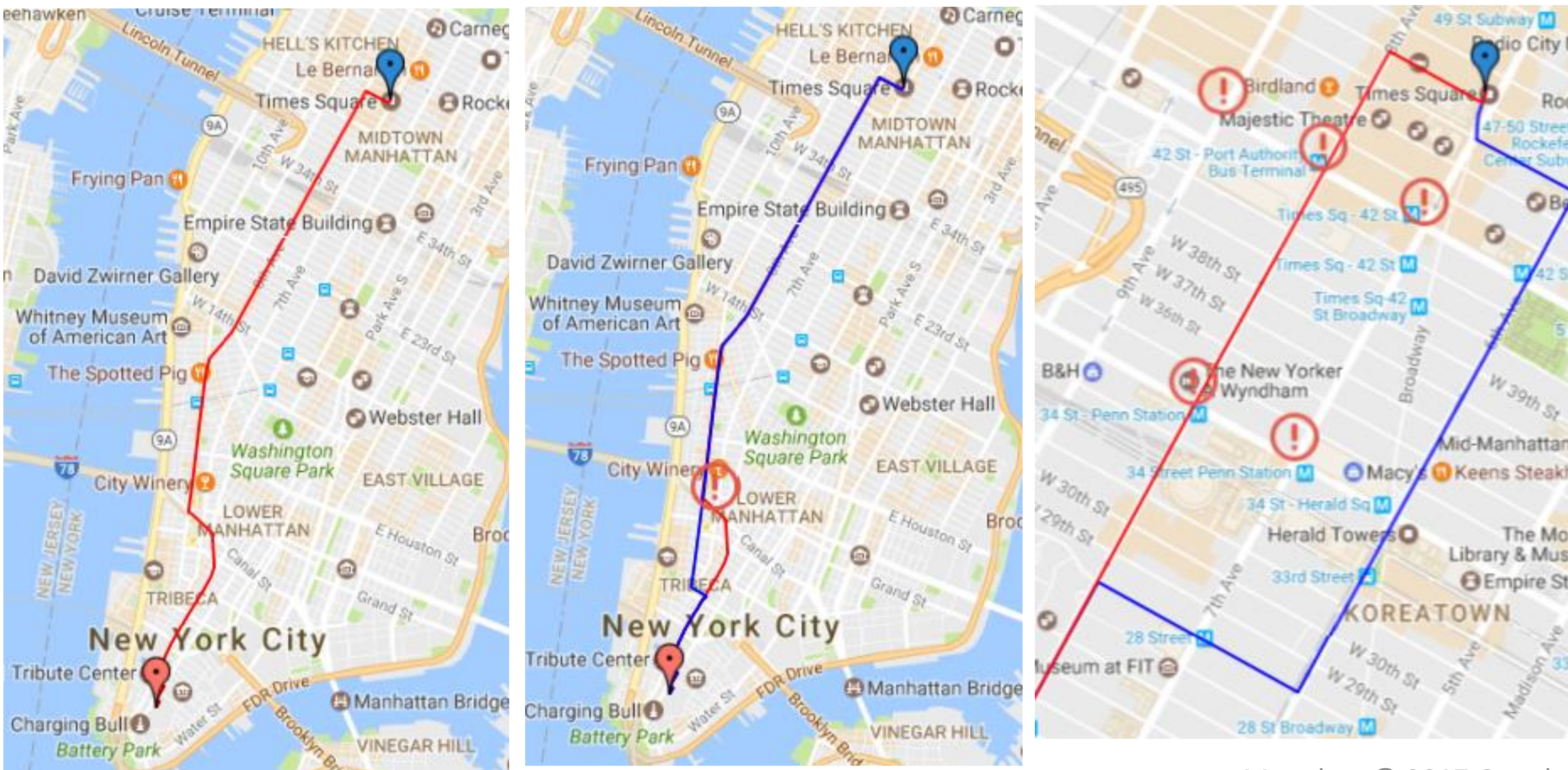
„How to travel safely in NYC?"

Workflow for a **safer** route:

1. Enter a start and an end location
2. Choose the means of transportation
3. Geoencode the entered adresses with Google API
4. Determine dangerous spots based on the dataset within a minimal bounding rectangle of the start and end locations while respecting a personal risk factor (lower values indicate a higher will to take risks)
5. Get the route via the HERE API while avoiding the dangerous spots
6. Plot the returned GPS coordinates on Google Maps

Example 1

- Let‘s assume we want to travel from “New York Stock Exchange” to “Times Square” with a bicycle.

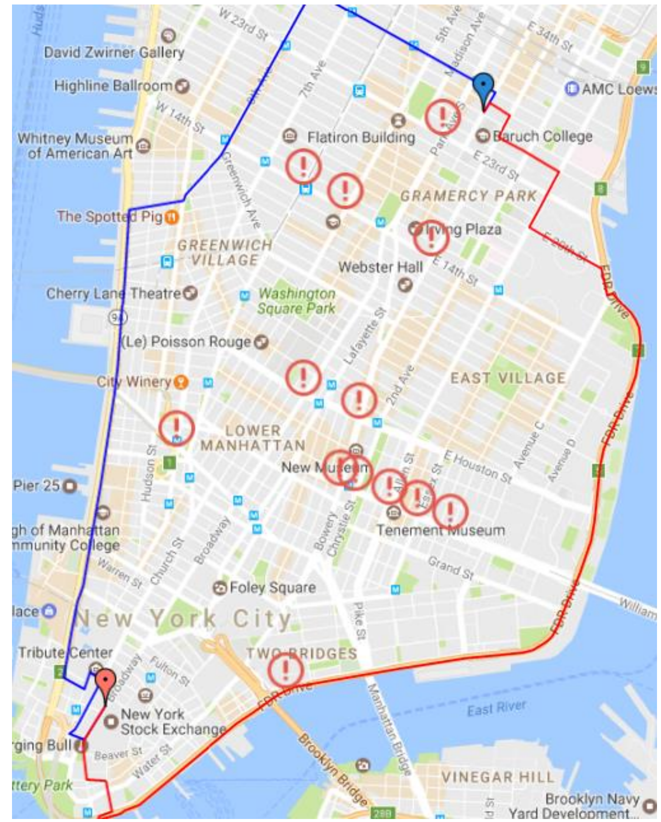


Map data © 2017 Google

Example 2

- Let‘s assume we want to travel from “New York Stock Exchange” to “e 86th street, Lexington avenue” with a car.

fastest route

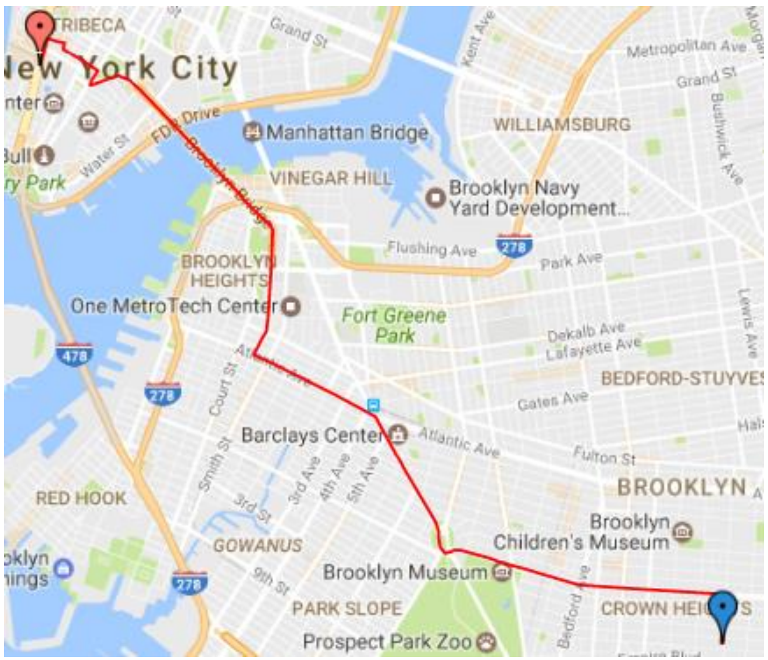


risk factor = 3

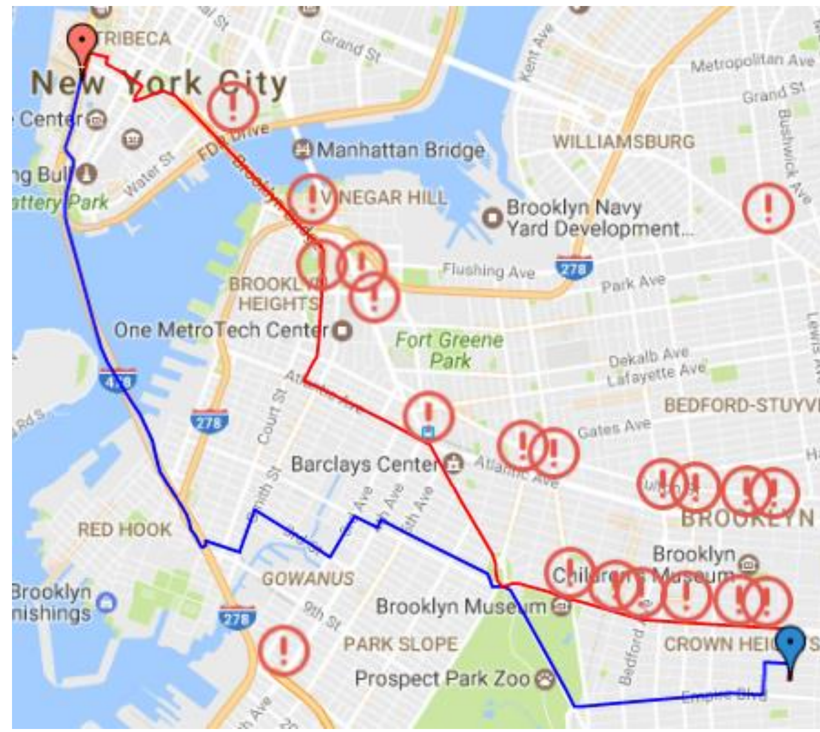
bit.do/hhu-btw2017

Example 3

- Another example that shows that our navigation works outside of Manhattan ;) “One World Trade Center” to “albany ave, prospect pl”



fastest route



risk factor = 5

bit.do/hhu-btw2017

Conclusion and Future Work

Conclusion & Future Work

Conclusion:

- Introducted the goal of the task
- Showed descriptive statistics in Azure ML
- Visualized the accidents in form of a heatmap
- Animated the development of the accidents over time
- Demonstrated a solution to possibly make transportation in NYC a little bit safer if our Azure-based solution is actually used by people before starting a trip

Future Work:

- Use Azure ML to predict the vehicle type code based on the features in the dataset
- Respect temporal aspects and weather conditions, e.g., the current season or events like Super Bowl in the route calculation

Thank you for your attention.