Maximizing The Effectiveness of Data

Abdulla Alhammadi |

2020-08-11

Contents

Chapter 1

Introduction

1.1 The Purpose

This project is a personal initiative of how setting better standards can help utilize modern technology to better collect and analyze data for specific applications, as well as how to make the most out of public data. In this case, the task will be to look into developing the standards for collecting public data, and how to make the most with what is available currently have to create better insights. In this case, for public policy.

1.2 Abstract

With the advancements in data technologies, the prevalence of data has skyrocketed in recent years. Due to this, the availability of data has increased significantly.

> "There are 2.5 quintillion bytes of data created every day— which is why you need IBM Cloud Object Storage"
>
> — an IBM Advertisement.

That is indeed an impressively large amount of data, but the IBM Cloud Object Storage is not actually needed because it is very likely that most of the data is probably random back-end logs or videos of cats. In any case, most of the data available may not be useful, usable data. This document will go through the importance of developing standards and criteria for the processes of gathering, entering, and handling of data for the purpose of developing high quality datasets to provide valuable and beneficial results.

Chapter 2

The Dataset

2.1 The Case

The case at hand is an example of public data that is available on Bayanat.ae. It is a dataset that includes data about traffic incidents that were reported to the Dubai Police operation contact center. This is the dataset that will be used to look for insights on how to improve road safety. The following pages will explain why this dataset was chosen, and how it is a good example for the ideas that will be mentioned in the following chapters.

2.2 Why this dataset?

This dataset was chosen specifically because it covers most of the ideas that will be brought up. Additionally, the dataset is a fair representation of the state of public data. This dataset provides a wide variety of types of observations (date and time, geographical data, nominal data, ordinal data, etc.) as well as a lot of room for improvement and utilization.

2.3 What will be covered?

The ideas that will be covered are:

- The 4 "key" aspects of data, which are:
  - Relevance
  - Reliability
  - Coverage

  - Consistency

- Setting the right standards
- Making the most of what is available

These are not the definitive and final points that make or break datasets, but more of a general guideline of best practices and a grasp of a positive thought process when working with such tasks. The coming chapters will go further into detail regarding the ideas listed above.

Chapter 3

The Key Aspects of Data

3.1 Relevance

Relevance in this context is essentially how fitting the data is with a task's set targets. For example in this case, the data is relevant because the task at hand is to explore the data to find any interesting relationships or occurrences that may help with providing solutions or insights to improve road safety. This is a glimpse of the dataset.

acci_id	acci_time	acci_name	acci_x	acci_y
3545008155	2019-06-10 11:59:50	-	25.28002	55.35302
3545009716	2019-06-10 12:04:19	-	25.25702	55.29077
3545011689	2019-06-10 12:09:48	-	25.17389	55.40356
3545013868	2019-06-10 12:18:18	-	25.26867	55.32277
3544995157	2019-06-10 11:22:00	-	25.26062	55.31896
3545003866	2019-06-10 11:47:55	-	25.08618	55.40152

In this case, the data is relevant because task provides enough flexibility to work with what is available and not asking certain questions or setting specific goals. To ensure that data is relevant, the entities that collect and provide the data should know what the users of the data are looking for. Collecting random data that may or may not be useful is not an effective way of providing data, since the efforts and resources used to gather and provide such data may be fruitless. A good example of ensuring that data is relevant is this hypothetical scenario:

> a person, X, is tasked with gathering data for the ABC Corp. annual performance report. Instead of collecting all the data that is possible, X decides to make the most out of his time by ensuring that the data that is intends to collect is relevant. So X meets with the committee that will be writing the report, and asks them what data is needed

to come up with the metrics required for a proper report that is valuable for stakeholders. By doing this, X has saved valuable time and effort as well as being much more productive than just blindly collecting data, in the hopes that the data may be relevant for the use case.

3.2 Reliability

Reliability is how much can the data be trusted to provide accurate and meaningful results. In this case for example, the data passed the initial reliability test, which tests the integrity of the data. This script is used to determine whether there were any inconsistencies between the cells.

```
check_doubles <- function(x){

  x <- x %>%
    group_by(acci_id) %>%
    filter(n_distinct(acci_name) > 1)

  if(nrow(x) > 1) {
    print(x)
  }
    else print("There are no inconsistencies between the cells")
}
```

This is a script that checks that each entry has a unique ID, which ensures that there are no duplicate entries or mismatched entries. However, the data proved to be unreliable because when the verifying the completeness of the data, this is what can be seen:

```
workfile %>% mutate(date = date(acci_time)) %>% count(is.na(date) == TRUE)
```

```
## # A tibble: 2 x 2
##   `is.na(date) == TRUE`     n
##   <lgl>                 <int>
## 1 FALSE                 10078
## 2 TRUE                  13743
```

It can be seen that more than half of the entries have no dates. This is a huge problem because dates can provide a lot of insights that is now very difficult to obtain. For example, a weather API could have been used to get the weather conditions on the dates, and study the effect that the weather conditions might have on road safety. There is another interesting observation.

```
workfile %>% mutate(incident_date = date(acci_time)) %>%
    mutate(incident_day = day(acci_time), Incident_month = month(acci_time)) %>%
    group_by(Incident_month) %>% count(incident_day) %>%
    count(Incident_month)
```

```
## # A tibble: 13 x 2
## # Groups:   Incident_month [13]
##    Incident_month     n
##             <dbl> <int>
##  1              1     3
##  2              2     3
##  3              3     3
##  4              4     4
##  5              5     3
##  6              6     4
##  7              7     3
##  8              8     3
##  9              9     3
## 10             10     3
## 11             11     3
## 12             12     3
## 13             NA     1
```

All the data collected is only from 3 to 4 days from each month. Whether this was an error or not, this can pose a lot of problems when exploring the data, and it is enough of a problem to make a lot of analysis useless. These issues when combined make the dataset unreliable. Yes, it is still possible to explore the data and possibly come up with conclusions, but these conclusions will have a very poor basis, making the results unreliable. A possible solution is automating the whole data collection process to reduce the possibility of collecting faulty data. Using a mobile app to submit incident reports is a solution that is currently being implemented by Dubai Police, and this solution is effective in ensuring the reliability of data because the data automatically registers the exact date and time of an incident.

3.3 Coverage

Coverage can be thought of as how much information the data ***can*** give. The amount of data that a dataset includes is affected by several factors, including the feasibility of data collection and relevance.

3.3.1 Using Technology to Cover More Data

In this case, modern technology could have been used to collect much more data, and the collected data could cover a lot of other variables such as the road features(bridge, junction, intersection, etc.), weather conditions (foggy, rainy, etc.), and much more. For example, Dubai Police receives traffic incidents that are submuitted via a mobile app. Meaning that the whole data collection process can be automated, including automatically obtaining the weather conditions on that day through the API's such as Accuweather, and the road features on the provided coordinates, which could be done through OpenStreetMap for example.

In this dataset, It is possible to increase the coverage by adding a new variable, the severity of the incidents. Because the severity of the incidents is entered with the accident type. So, by splitting the `acci_type` into two columns: `Incident_type` and `severity`, it is possible to determine the severity of the incidents and further explore the variable.

```
workfile <- workfile %>% separate(acci_name, c("Incident_type", "severity"), sep = "-
kable(head(workfile))
```

acci_id	acci_time	Incident_type	severity	acci_x	acci_y
3545008155	2019-06-10 11:59:50			25.28002	55.35302
3545009716	2019-06-10 12:04:19			25.25702	55.29077
3545011689	2019-06-10 12:09:48			25.17389	55.40356
3545013868	2019-06-10 12:18:18			25.26867	55.32277
3544995157	2019-06-10 11:22:00			25.26062	55.31896
3545003866	2019-06-10 11:47:55			25.08618	55.40152

Now, it is possible to explore the relationships between the severity of the incidents and other variables. By increasing the coverage of the data, there are much more insights that can be gathered than before.

3.4 Consistency

This is a very simple aspect, but is quite often overlooked. Having specific guidelines and standards to follow when collecting and using data is important to have consistent results. For example, if the data is collected but with different variables next year, it can be difficult to track and compare changes between now and then. Therefore, a solid framework that allows for both consistent results and the improvement of data quality is important to maximize the usefulness of the data.

Chapter 4

Foundations

4.1 Standards

Setting standards for data collection is vital for providing equal access to a wide variety of users as well as a wide variety of applications. Specific standards that are to be followed in the data collection process are essential to creating a robust foundation that makes data much more impactful. For example, setting standards for the key aspects of data can make sure that the results are consistent as well as reproducible for many other applications. In this dataset, setting standards for data entry could have made the dataset less difficult to work with, as it can be seen from the instance where one column was split into the two columns `Incident_type` and `severity` from the previous chapter because the two scopes of data were coerced into one variable, which had to be adjusted to become useful. Another standard, such as adding indexes, is very helpful because it makes identifying nominal data much easier, as well as adding a common key for future purposes, such as translation. In this case, an ID for severity was added. This would allow a user to manipulate the data later on.

```
library(tidyverse)
library(knitr)
dfcount <- workfile %>% count(severity)
indexlegend <- dfcount %>% dplyr::select(severity) %>% na.omit() %>% mutate(severity_ID = c(2, 3,
workfile <- workfile %>% left_join(indexlegend)
kable(head(workfile))
```

acci_id	acci_time	Incident_type	severity	acci_x	acci_y	severity_ID
3545008155	2019-06-10 11:59:50			25.28002	55.35302	3
3545009716	2019-06-10 12:04:19			25.25702	55.29077	1
3545011689	2019-06-10 12:09:48			25.17389	55.40356	1
3545013868	2019-06-10 12:18:18			25.26867	55.32277	1
3544995157	2019-06-10 11:22:00			25.26062	55.31896	1
3545003866	2019-06-10 11:47:55			25.08618	55.40152	1

4.2 Making the Most out of Things

Sometimes, there is only so much that can be done regarding problems in datasets. In such circumstances, it is important to make the most of what is available. In this case, using the packages `httr` and `JSONlite` to import data from various API's such as OpenStreetMap and AccuWeather, and add the data to new variables such as `weather_condition` and `road_feature`. The given coordinates can be used also to add road features and assign it to another variable, or use the dates provided to get the weather on a given date and add it to the data. By doing these things, it has increased the dataset's coverage massively, making it significantly more valuable.

Chapter 5

Final Words

Thank you for reading this document. All of the code used here is my own personal work. I hope I was able to help you better understand about how important quality data is, and how better data and some creativity can help improve insights significantly.