Data Warehouses – Report04

Wroclaw University of Science and Technology, Date: April 6, 2022

Student:	----------	Grade
Identifier	245784	
First name	Rahul	
Last name	Vijayvargiya	

This laboratory assignment consists of 2 tasks. If you cannot solve the task, try to give at least a partial solution or justification for the reason for the lack of a solution.

Data Source: AviationData.csv

Task 1

Study the provided dataset and present your finding in the following scope:

- *Domain data dictionary – with information about attribute name, attribute type (high-level type representation, like numerical, money, text, date, etc.), description (short description of the meaning of the attribute).*
- *Quality assessment of source data – with information about table/sheet/location, attribute name, attribute type (lower level type representation, like varchar(20), decimal(5,2), etc.), type of data (nominal, ordinal, interval, ratio, continuous), number of unique values, null ratio, quality assessment description (short description of the results of the attribute quality assessment – focusing on a column consistency assessment).*

In the resultant tables, please mark all occurrences of questionable (in terms of further usage in data analysis) attributes; please remember to later justify your selection and decisions.

Use this section to provide your solutions; please remember to present two tables:

Solution:

1 – Domain Data Dictionary

Interpretation of data:

File: AviationData.csv			
	Attribute	Value type	Meaning
1.	Event Id	Text	Code identifying the date and number of the incident, provided for each incident
2.	Investigation Type	Text	Type of investigation
3	Accident Number	Text	Code identifying the accident number
4	Event Date	Date	Date of the accident event
5	Location	Text	Place (location) of the accident. City, place and abbreviation of the province.

6	Country	Text	The country where the accident was recorded.
7	Latitude	Numeric	Coordinate of the severity of the accident.
8	Longitude	Numeric	Coordinate of the longitude of the accident.
9	Airport Code	Text	The code identifying the airport.
10	Airport Name	Text	The name of the airport.
11	Injury Severity	Text	Severity of injuries.
12	Aircraft Damage	Text	Damage to the aircraft.
13	Aircraft Category	Text	Aircraft category (aircraft, helicopter, ...).
14	Registration Number	Text	Registration number of the transport.
15	Make	Text	Brand (construction) of transport.
16	Model	Text	Ship model.
17	Amateur Built	Text	Was it built amateur.
18	Number of Engines	Numeric	Number of engines.
19	Engine Type	Text	Engine type.
20	FAR Description	Text	Description of Federal Aviation Charts
21	Schedule	Text	Flight schedule
22	Purpose of Flight	Text	Destination of the flight.
23	Air Carrier	Text	Aviation trigger.
24	Total Fatal Injuries	Numeric	Total fatal injuries.

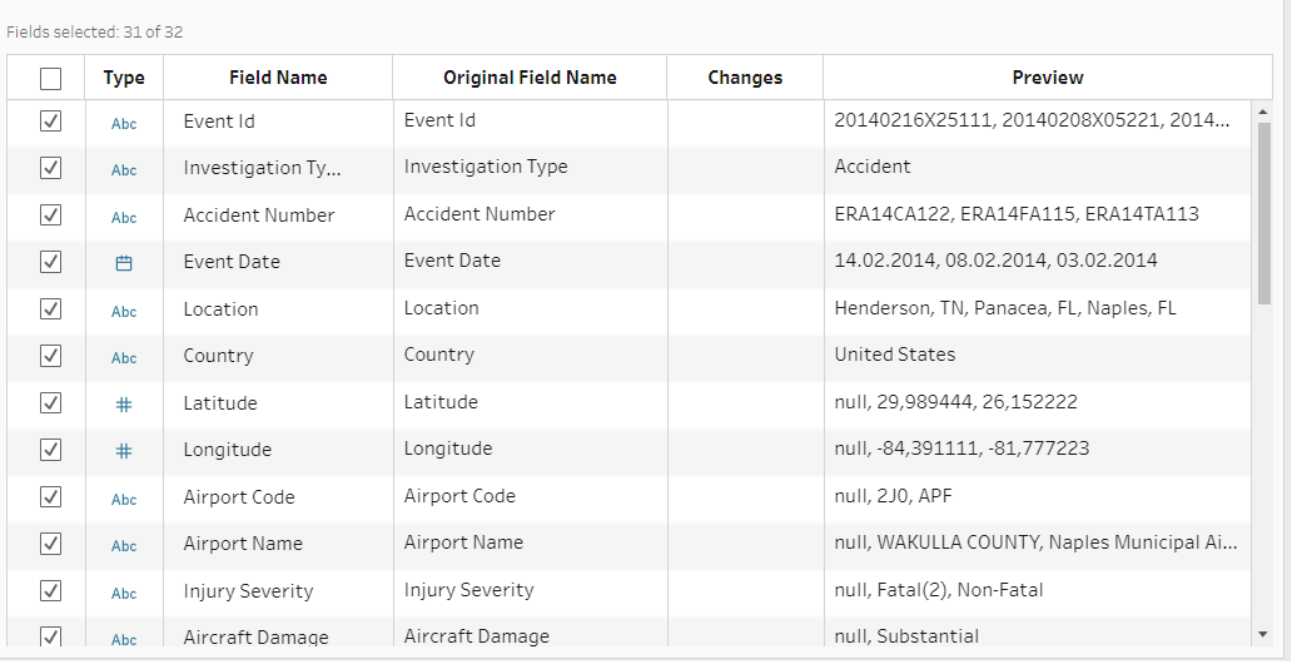
25	Total Serious Injuries	Numeric	The total number of serious injuries.
26	Total Minor Injuries	Numeric	The total number of minor injuries.
27	Total Uninjured	Numeric	Amount without injury.
28	Weather Condition	Text	The state of the weather at the moment of flight.
29	Broad Phase of Flight	Text	Wide flight phase.
30	Report Status	Text	Report status.
31	Publication Date	Date	The date of publication of the accident.

2 – Quality assessment Sheet:

File: AviationData.csv				Number of records: 49,997
	Attribute	Type	Value range	Data quality assessment
1.	Event Id	Text	Length: 14	0% null, 49289 unique
2.	Investigation Type	Text	Accident (96%) or Incident (4%)	One meaning null (<1%), 3 unique
3	Accident Number	Text	Length: 9-11	0% null, 49997 unique
4	Event Date	Date	01.01.1980 - 01.01. 2020	8708 unique meanings
5	Location	Text	Length: 4-61	<1% null, 18965 unique
6	Country	Text	Length: 4-30	<1% null, 163 unique
7	Latitude	Numeric	-80 – 90	57% null, 14810 unique
8	Longitude	Numeric	-200 – 200	57% null, 15726 unique
9	Airport Code	Text	Length: 1-8	42% null 7563 unique
10	Airport Name	Text	Length: 2-33	40% null, 16352 unique

11	Injury Severity	Text	Length: 8-11	<1% null, 105 unique
12	Aircraft Damage	Text	"Destroyed" (20%), "Minor" (3%), or "Substantial" (74%)	3% null, 4 unique
13	Aircraft Category	Text	Length: 5-12	Ship category, 79% null
14	Registration Number	Text	Length: 3-11	5% null, 44305 unique
15	Make	Text	Length: 2-30	<1% null, 5805 unique
16	Model	Text	Length: 1-20	<1% null, 8556 unique
17	Amateur Built	Text	"Yes" (11%) or "No" (88%)	1% null, 3 unique
18	Number of Engines	Numeric	0-4	6% null, 6 unique
19	Engine Type	Text	Length: 4-16	6% null, 16 unique
20	FAR Description	Text	Length: 7-30	79% null,
21	Schedule	Text	"NSCH" (5%), "SCHD" (5%), "UNK" (4%)	86% null
22	Purpose of Flight	Text	Length: 4-19	6% null, 23 unique
23	Air Carrier	Text	Length: 3-90	95% null, very unhelpable data
24	Total Fatal Injuries	Numeric	0-350	39% null
25	Total Serious Injuries	Numeric	0-110	42% null
26	Total Minor Injuries	Numeric	0-375	40% null
27	Total Uninjured	Numeric	0-700	20% null
28	Weather Condition	Text	"IMC" (7%), "UNK" (<1%), "VMC" (89%)	3% null
29	Broad Phase of Flight	Text	<1% "UNKNOWN", <1% "OTHER"	12% null, 13 unique groups

30	Report Status	Text	Length: 7-14	0% null
31	Publication Date	Date	01.01.1990 – 01.01. 2020	< 1% null



Fields selected: 31 of 32

Type	Field Name	Original Field Name	Changes	Preview
Abc	Event Id	Event Id		20140216X25111, 20140208X05221, 2014...
Abc	Investigation Ty...	Investigation Type		Accident
Abc	Accident Number	Accident Number		ERA14CA122, ERA14FA115, ERA14TA113
📅	Event Date	Event Date		14.02.2014, 08.02.2014, 03.02.2014
Abc	Location	Location		Henderson, TN, Panacea, FL, Naples, FL
Abc	Country	Country		United States
#	Latitude	Latitude		null, 29,989444, 26,152222
#	Longitude	Longitude		null, -84,391111, -81,777223
Abc	Airport Code	Airport Code		null, 2J0, APF
Abc	Airport Name	Airport Name		null, WAKULLA COUNTY, Naples Municipal Ai...
Abc	Injury Severity	Injury Severity		null, Fatal(2), Non-Fatal
Abc	Aircraft Damage	Aircraft Damage		null, Substantial

Conclusions

Use this section to provide insights on your methodology, i.e., the process you have utilized to solve the task.

We imported Data into Tableau prep, Did bit Data Cleaning and Analyzing and generated a fix version of provided dataset, Aviation Dataset for further use

---

Task 2

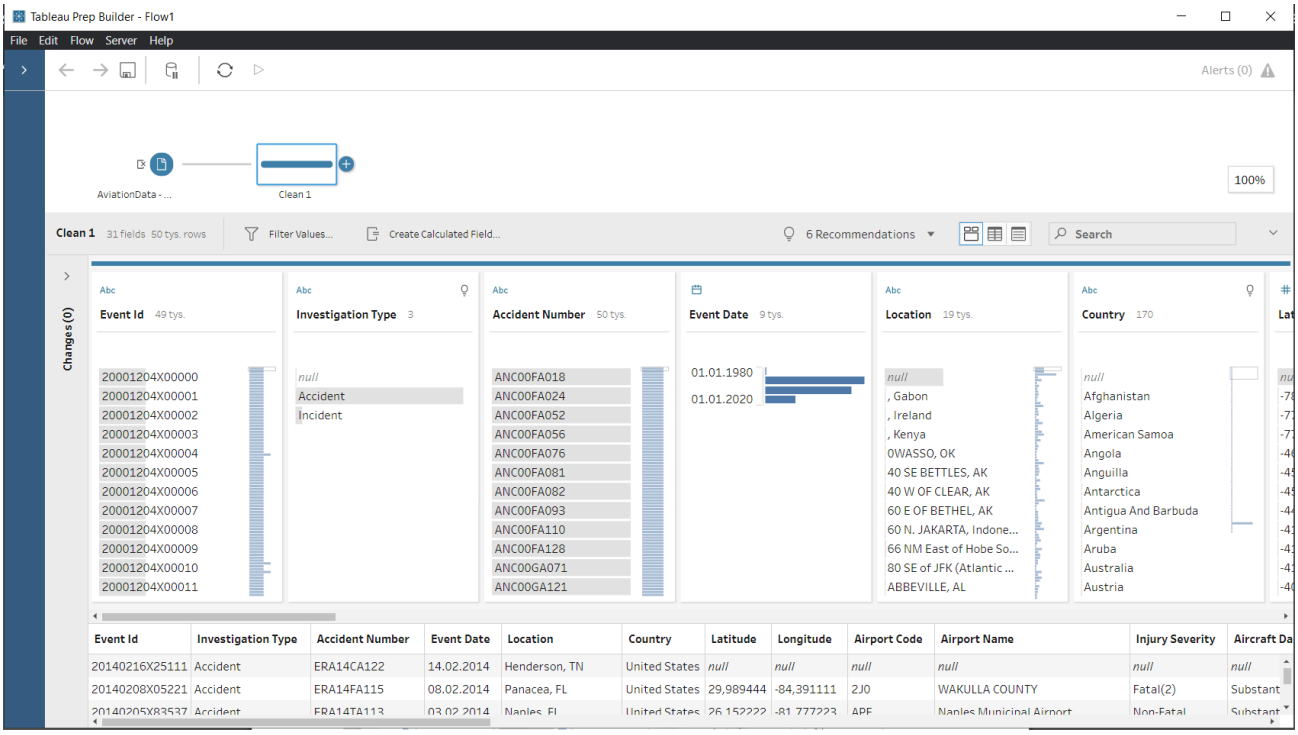
Let us now look at some data wrangling capabilities of the Tableau Prep tool. In particular, try cleaning some of the quality issues present in the "AviationData.xls" dataset.

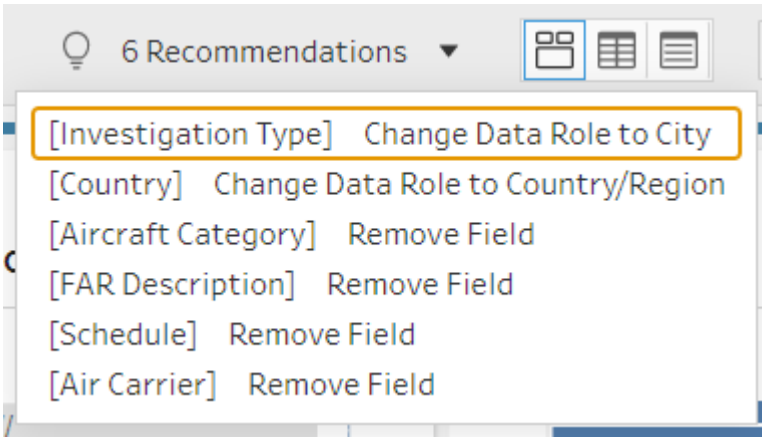
Prepare a flow in Tableau Prep to clean 4 (selected examples) of the identified (in the previous task) types of data quality issues. Resultant data, please store in a local SQL Server database table.

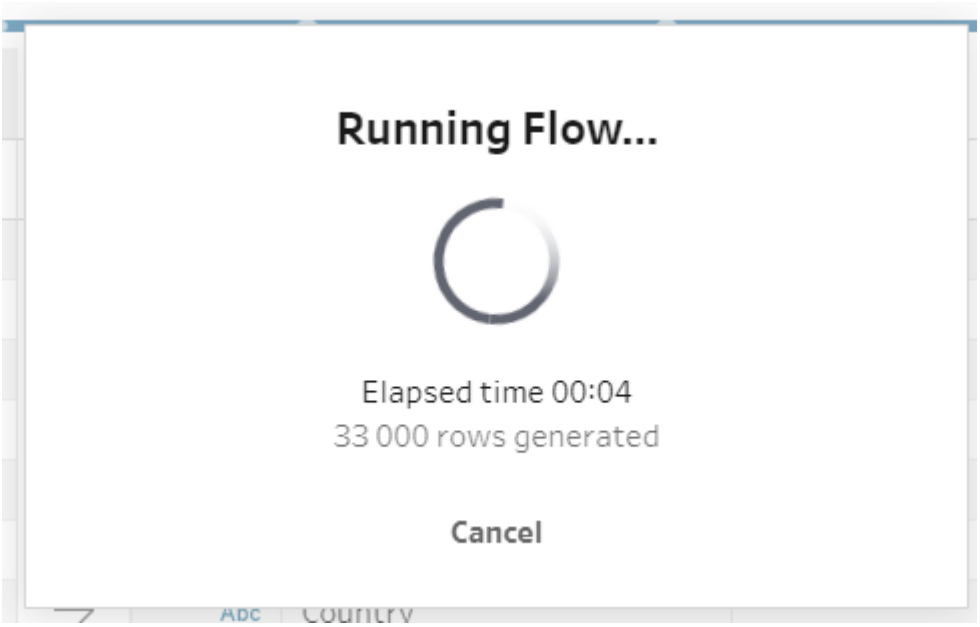
Use this section to provide your solutions; please remember to present a screenshot of the created flow and short description (how the quality issues were handled):

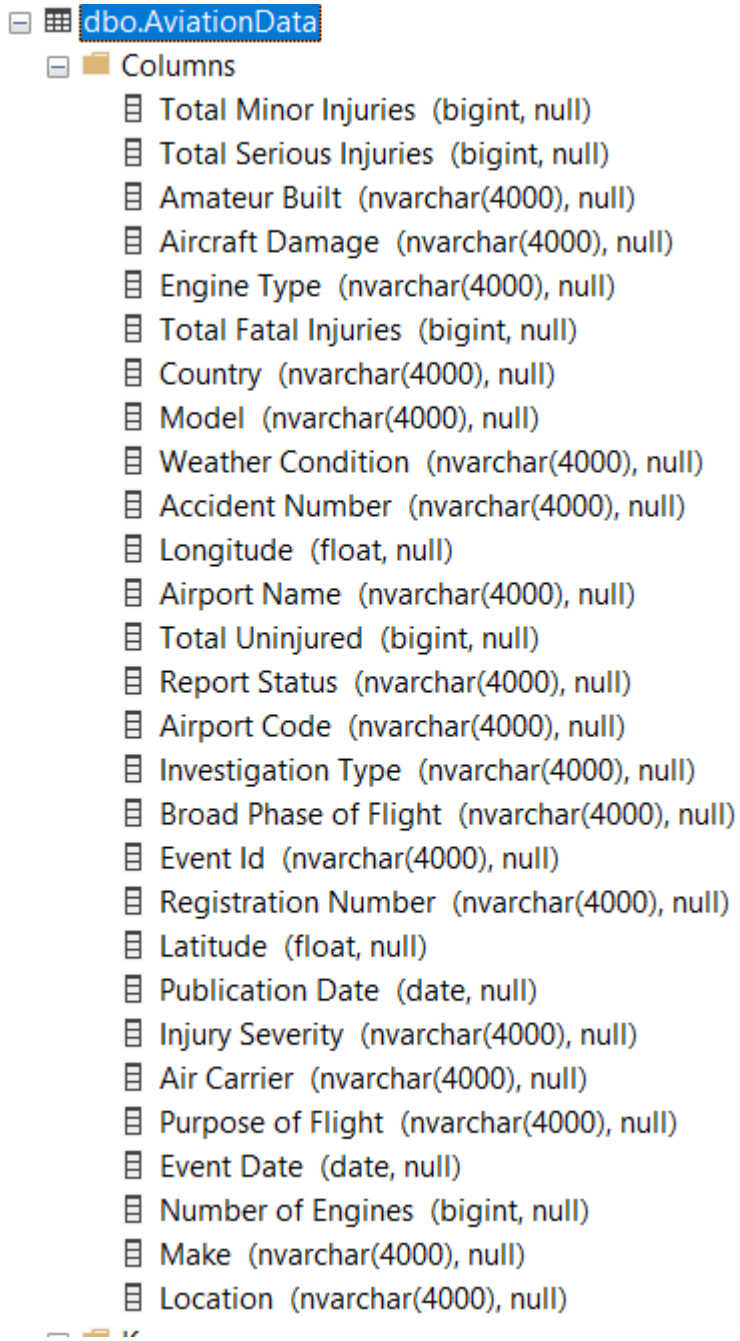
Solution:

Flow (screenShot)






Running Flow...
Elapsed time 00:04
33 000 rows generated
Cancel
Abc
Country



	Total Minor Injuries	Total Serious Injuries	Amateur Built	Aircraft Damage	Engine Type	Total Fatal Injuries	Country	Model	Weather Condition	Accident Number	Longitude	Airport N
31...	0	0	No	Substantial	Reciprocating	0	United States	PA-32RT-300	VMC	FTW98LA106	NULL	TAOS
31...	0	1	No	Substantial	Reciprocating	0	United States	PA-38-112	VMC	ATL98LA040	NULL	LAUREI
31...	0	0	No	Destroyed	Reciprocating	2	NULL	G35	UNK	MIA98FAMS1	NULL	NULL
31...	0	0	No	Destroyed	Reciprocating	2	United States	M20J	VMC	LAX98FA050	NULL	NULL
31...	NULL	NULL	Yes	Minor	Unknown	NULL	France	MD83-2	UNK	DCA98WA017	NULL	NULL
31...	0	0	No	Substantial	Reciprocating	0	United States	140A	VMC	FTW98LA105	NULL	KEN W
31...	0	0	No	Substantial	Reciprocating	0	United States	340A	VMC	LAX98LA081	NULL	PINE M
31...	1	0	No	Substantial	Reciprocating	0	United States	269C	VMC	FTW98TA104	NULL	NULL
31...	0	0	No	Substantial	Reciprocating	0	United States	182Q	VMC	FTW98LA113	NULL	PRIVATI
31...	0	0	No	Substantial	Turbo Prop	0	United States	65-A90	IMC	ATL98LA038	NULL	ROBER
31...	0	0	No	NULL	Reciprocating	0	United States	R-44	VMC	ATL98IA039	NULL	KENDAI
31...	0	0	No	Substantial	Reciprocating	0	United States	150	VMC	ANC98LA016	NULL	NULL
31...	1	0	No	Substantial	Turbo Prop	0	United States	ATR-42-320	VMC	NYC98FA062	NULL	BRADLI
31...	0	0	No	Destroyed	Turbo Prop	3	United States	695A	VMC	ATL98FA036	NULL	BOCA R
31...	0	0	No	Substantial	Turbo Prop	0	United States	1900D	IMC	IAD98LA023	NULL	SARAN/
31...	0	0	Yes	Substantial	Reciprocating	0	United States	CHALLENGE...	VMC	FTW98LA408	NULL	PRIVATI

Column Null Ratio Profiles - [dbo].[AviationData]

Column	Null Count	Null Percentage
Accident Number	0	0.0000 %
Air Carrier	47522	95.1068 %
Aircraft Damage	1633	3.2682 %
Airport Code	20988	42.0037 %
Airport Name	19788	39.6021 %
Amateur Built	516	1.0327 %
Broad Phase of Flight	5883	11.7738 %
Country	285	0.5704 %
Engine Type	2750	5.5036 %
Event Date	0	0.0000 %
Event Id	0	0.0000 %
Injury Severity	70	0.1401 %
Investigation Type	1	0.0020 %
Latitude	28505	57.0477 %
Location	79	0.1581 %
Longitude	28514	57.0657 %
Make	82	0.1641 %
Model	104	0.2081 %
Number of Engines	2977	5.9579 %
Publication Date	392	0.7845 %
Purpose of Flight	3035	6.0740 %
Registration Number	2329	4.6611 %
Report Status	0	0.0000 %
Total Fatal Injuries	19333	38.6915 %
Total Minor Injuries	20208	40.4427 %
Total Serious Injuries	21142	42.3119 %
Total Uninjured	10097	20.2073 %
Weather Condition	1544	3.0900 %

Column Statistics Profiles - [dbo].[AviationData]

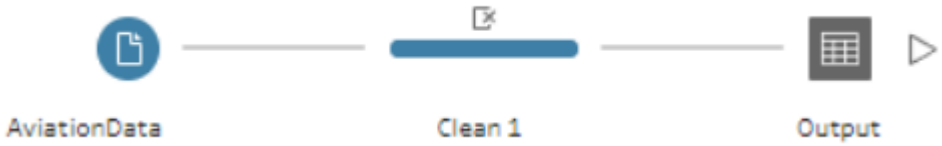
Column	Minimum	Maximum	Mean	Standard Deviation
Event Date	18.12.1989 00:0...	14.02.2014 00:0...		
Latitude	-78,016945	89,218056	37.897895721694	11.788400559987
Longitude	-193,216667	177,557778	-96.1690762400...	34.0074975095298
Number of Engines	0	4	1.1438603958289	0.442321251433874
Publication Date	26.04.1990 00:0...	14.02.2014 00:0...		
Total Fatal Injuries	0	349	1.01269830906...	7.07506469630744
Total Minor Injuries	0	380	0.59850801438...	3.39280193970181
Total Serious Injuries	0	111	0.38300086730...	1.70901056187677
Total Uninjured	0	699	6.38234261349...	30.6849708063819

Column Value Distribution Profiles - [dbo].[AviationData]

Column	Number Of Distinct Values
Accident Number	49967
Air Carrier	1825
Aircraft Damage	3
Airport Code	7563
Airport Name	16351
Amateur Built	2
Broad Phase of Flight	12
Country	162
Engine Type	15
Event Date	8708
Event Id	49259
Injury Severity	104
Investigation Type	2
Location	18939
Make	5804
Model	8556
Number of Engines	5
Publication Date	2596
Purpose of Flight	22
Registration Number	44304
Report Status	4
Total Fatal Injuries	102
Total Minor Injuries	55
Total Serious Injuries	38
Total Uninjured	331
Weather Condition	3

FLOW (Description)

We removed all Null Values, Process Dataset with the help of tableau prep and later we insert data into database



CONCLUSIONS:

Use this section to provide your conclusions:

For the Above Task we use Tableau prep because it was easy, user-friendly app, things we did in Tableau could be done as well in Python, Python also has great of Packages for Data analysis and File Handling,

From this Lab we learned Basic Data Analysis with Tableau Prep, Handling CSV files and How Use Tableau Prep

---

REMARKS:

- *A report without final conclusions will not be checked and results in a negative score!*
- *The report file should be named **Rep04DW-StudentID-Last name-2022**, please use the PDF format*
- *You should use MS SQL SERVER 2019 (or 2017), Visual Studio and Tableau Prep (available at https://www.tableau.com/academic/students)*