ETL-Prozess anhand von Inlandflügen der USA (2015)

CIP-Projekt Gruppe 16

Hochschule Luzern – Wirtschaft

MSC Applied Information and Data Science

Data Collection, Integration, Preprossessing – Projektarbeit (LN2)

Agenda

Aufgaben-/ Fragestellung

Datenquellen

Tools

ETL-Prozess

Python

Tableau Prep

SQL Server

Analyse mit Tableau Desktop

Fazit und Reflexion

Aufgaben-/ Fragestellung

Aufgabenstellung:

Generieren eines **ETL**-Prozesses:

- **E**xtract:
  - Datenbeschaffung aus mind. zwei Quellen
- **T**ransform:
  - Aufbereitung und Bereinigung der Daten
  - Verknüpfung der Datensätze
- **L**oad:
  - Daten in Datenbank importieren

Fragestellung:

Welche der Airlines haben die meisten pünktlichsten, verspäteten, annullierten und umgeleiteten Flüge verbucht.

Datenquellen

- Inlandflüge 2015 der USA
  - flights.csv
  - airlines.csv
  - airports.csv

kaggle

- Rufzeichen der Airlines
  - Gecrawlt mit Python

Software / Verwendete Tools

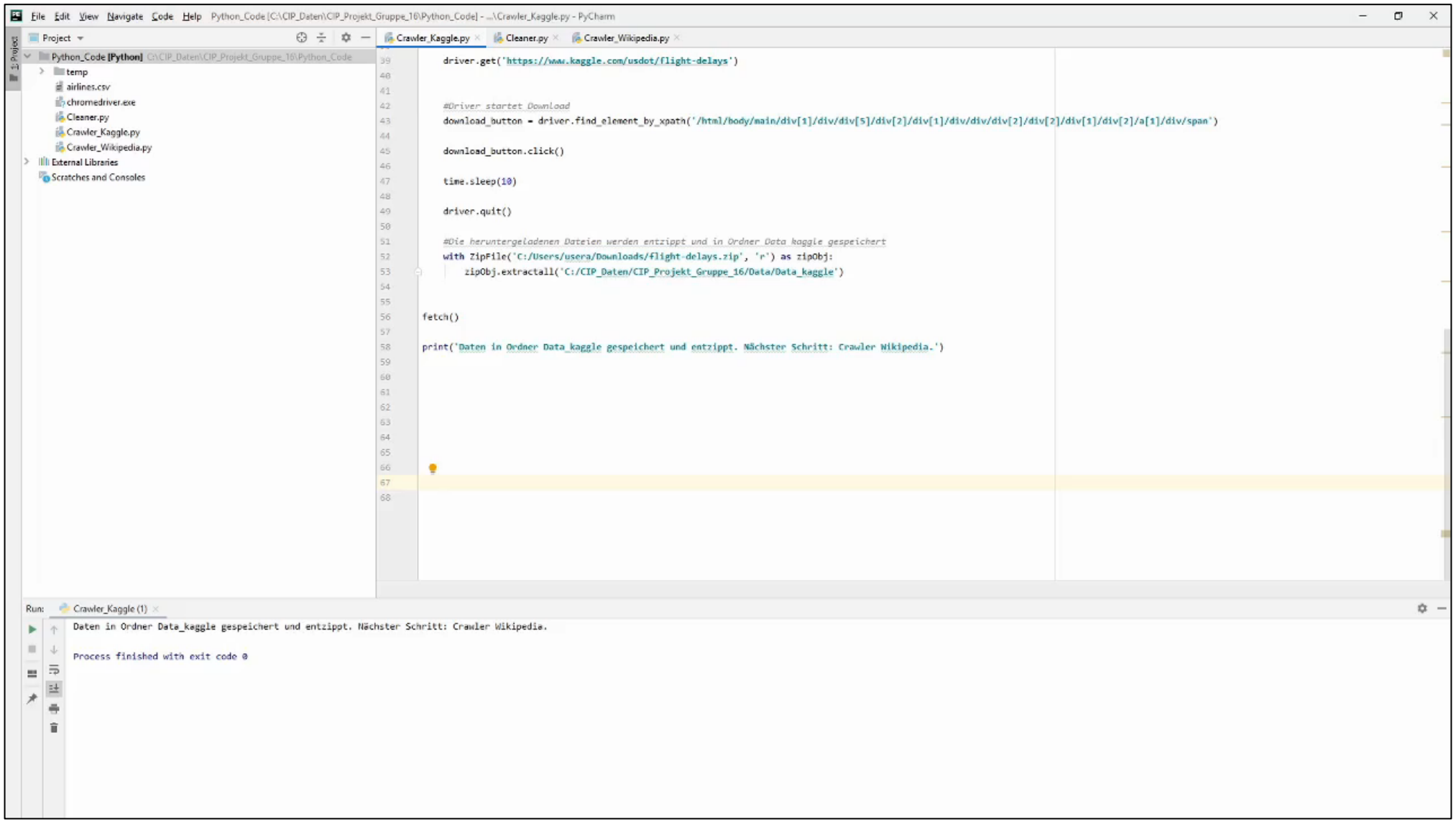
Extract
- Python – Selenium – BeautifulSoup

Transform
- Python – Pandas - Numpy
- Tableau Prep

Load
- Microsoft SQL-Server

Visualise
- Tableau Desktop

Crawler


File Edit View Navigate Code Help Python_Code [C:\CIP_Daten\CIP_Projekt_Gruppe_16\Python_Code] - ...\Crawler_Kaggle.py - PyCharm
Project
Python_Code [Python] C:\CIP_Daten\CIP_Projekt_Gruppe_16\Python_Code
temp
airlines.csv
chromedriver.exe
Cleaner.py
Crawler_Kaggle.py
Crawler_Wikipedia.py
External Libraries
Scratches and Consoles
Crawler_Kaggle.py
Cleaner.py
Crawler_Wikipedia.py
driver.get('https://www.kaggle.com/usdot/flight-delays')
#Driver startet Download
download_button = driver.find_element_by_xpath('/html/body/main/div[1]/div/div[5]/div[2]/div[1]/div/div/div[2]/div[2]/div[1]/div[2]/a[1]/div/span')
download_button.click()
time.sleep(10)
driver.quit()
#Die heruntergeladenen Dateien werden entzippt und in Ordner Data kaggle gespeichert
with ZipFile('C:/Users/usera/Downloads/flight-delays.zip', 'r') as zipObj:
zipObj.extractall('C:/CIP_Daten/CIP_Projekt_Gruppe_16/Data/Data_kaggle')
fetch()
print('Daten in Ordner Data_kaggle gespeichert und entzippt. Nächster Schritt: Crawler Wikipedia.')
Run:
Crawler_Kaggle (1)
Daten in Ordner Data_kaggle gespeichert und entzippt. Nächster Schritt: Crawler Wikipedia.
Process finished with exit code 0

Datenbereinigung mit Python

- Analyse der Daten
- Aufbereitung und Bereinigung der Daten

Herausforderungen:

- Datenmenge
- Unvollständige Datensätze
- Nicht eindeutig zuweisbare Einträge

In [4]:
```
print(len(df_main))
```

```
5819079
```

In [5]:
```
df.head()
```

Out[5]:

	YEAR	MONTH	DAY	DAY_OF_WEEK	AIRLINE	FLIGHT_NUMBER	TAIL_NUMBER	ORIGIN_AIRPORT	DESTINATION_AIRPORT	SCHEDULED
0	2015	1	1	4	AS	98	N407AS	ANC	SEA	
1	2015	1	1	4	AA	2336	N3KUAA	LAX	PBI	
2	2015	1	1	4	US	840	N171US	SFO	CLT	
3	2015	1	1	4	AA	258	N3HYAA	LAX	MIA	
4	2015	1	1	4	AS	135	N527AS	SEA	ANC	

5 rows × 31 columns

In [7]:
```
# TESTING
is_october = df['MONTH'] == 10

a = df[is_october]
a.head()
```

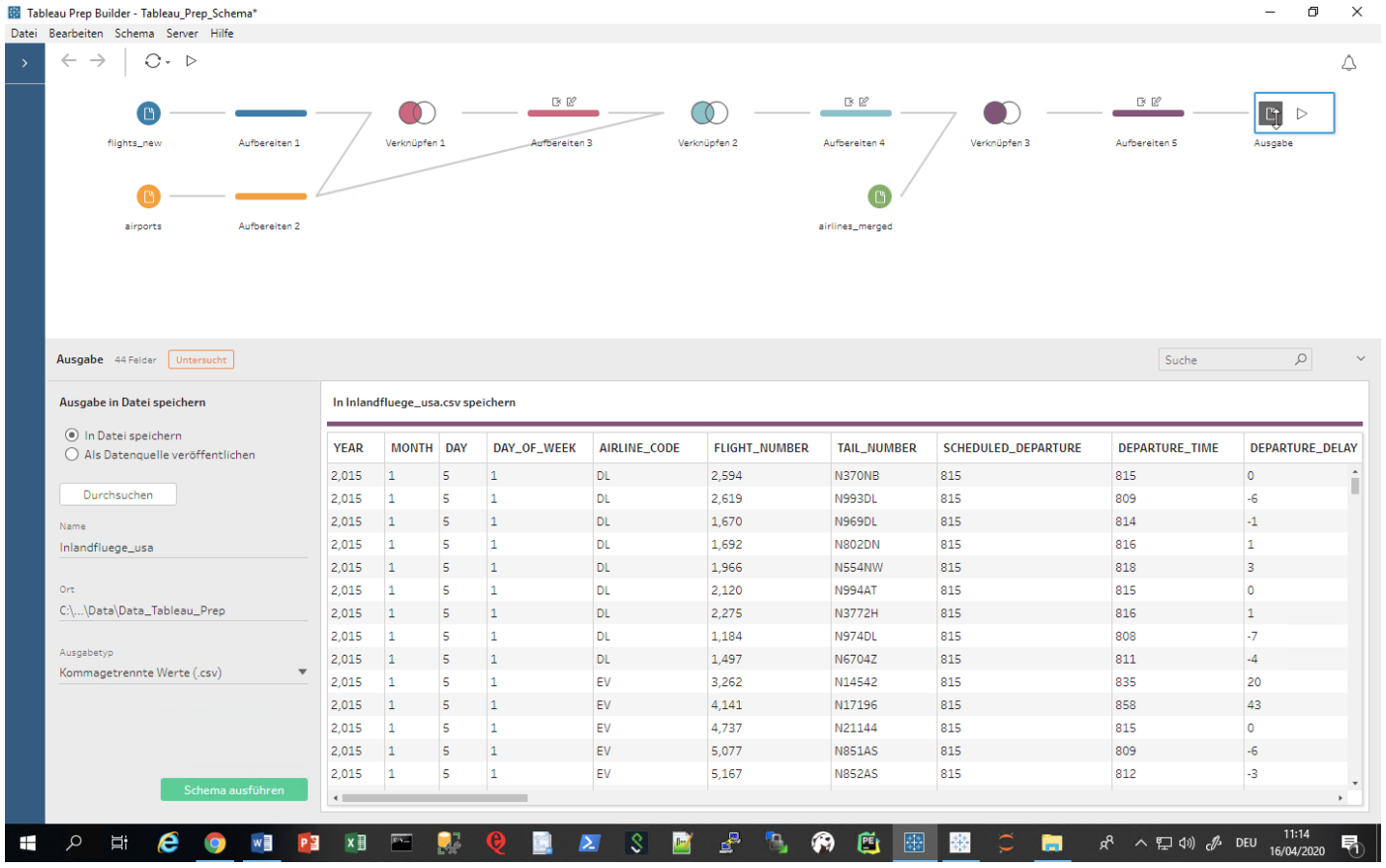
Out[7]:

	YEAR	MONTH	DAY	DAY_OF_WEEK	AIRLINE	FLIGHT_NUMBER	TAIL_NUMBER	ORIGIN_AIRPORT	DESTINATION_AIRPORT	SCHE
4385712	2015	10	1	4	AA	1230	N3DBAA	14747	11298	
4385713	2015	10	1	4	DL	1805	N696DL	14771	13487	

Transform

- **Zusammenführen der Dateien und Anpassen der Spalten zu einer Ausgabedatei**

- **Herausforderungen:**
  - **Die richtige Auswahl der Verknüpfung (left, inner oder right join)**
  - **Anpassung der Spalten nach jedem join**

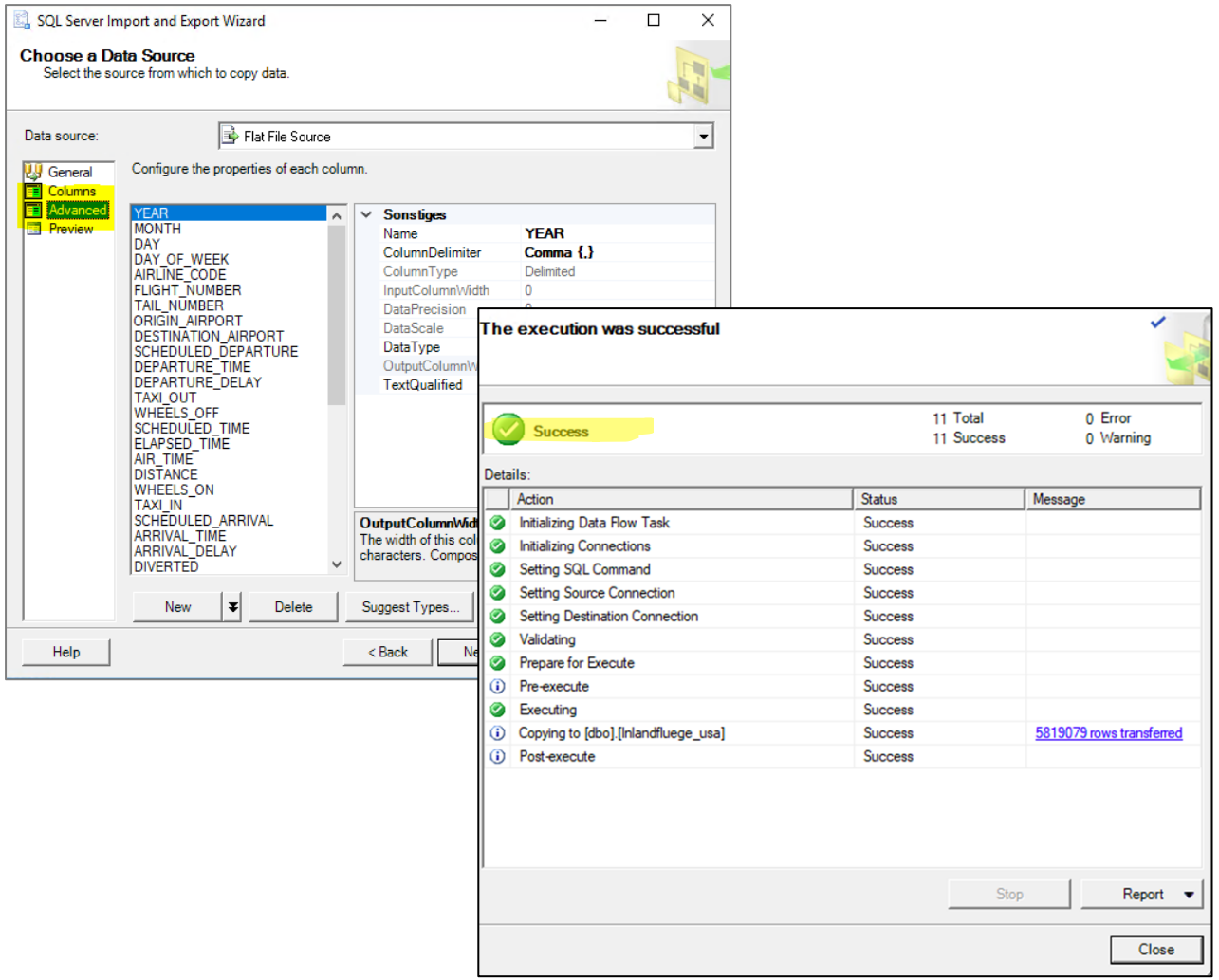


Load | SQL-Server Import

- **Erstellung einer Datenbank CIP:**
- **Import des CSV-Files via Wizard**

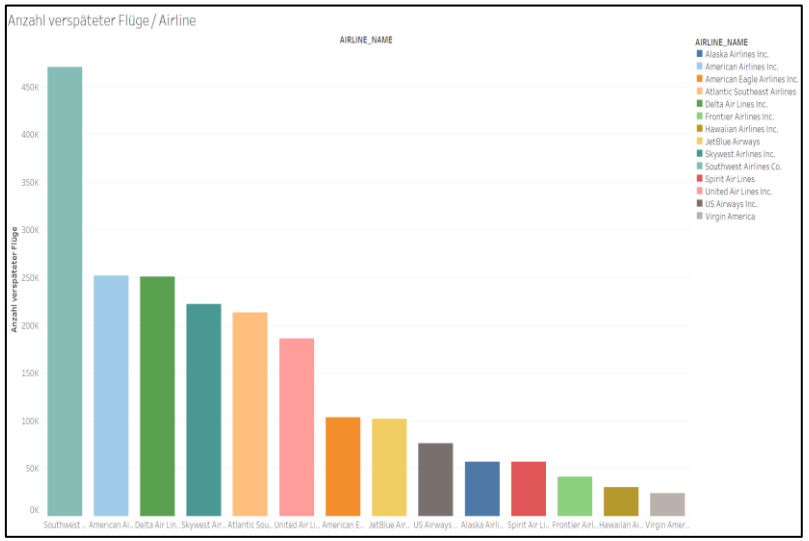
- **Herausforderungen:**
  - **Wenige Kenntnisse**
  - **Anpassung der Datentypen und Feldlänge**
  - **Dauer des Ladeprozesses**

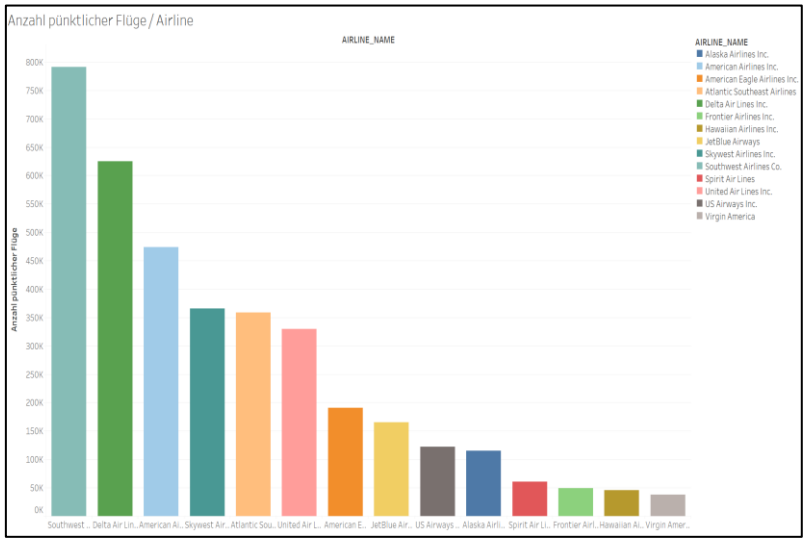
- **Erfolgreicher Import**

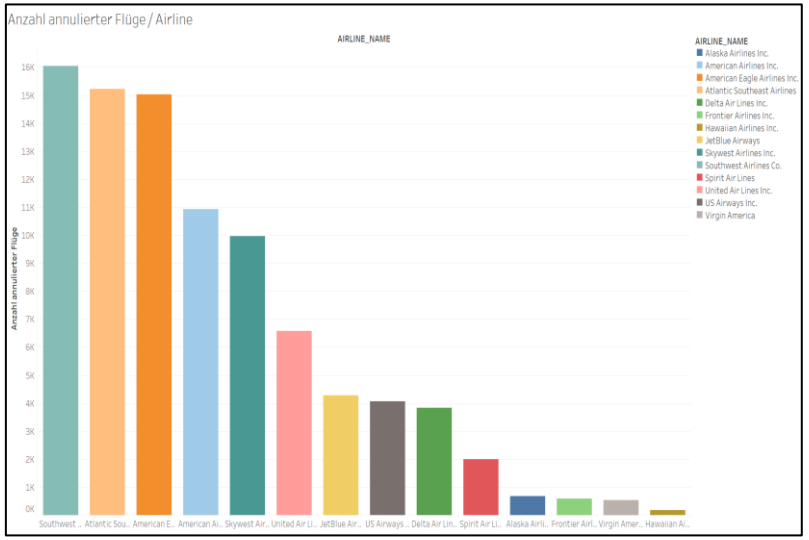


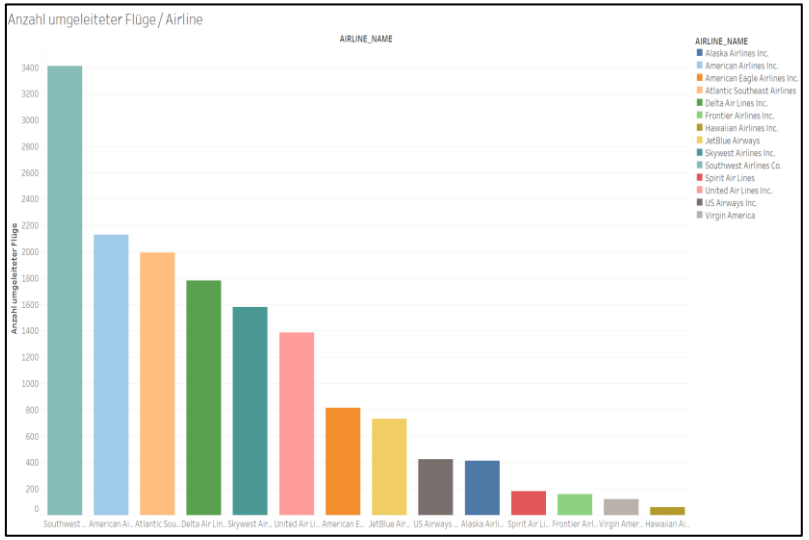
Analyse mit Tableau Desktop absolute Zahlen

- Grafiken sehr ähnlich
  - Southwest Airlines Co. immer führend
  - Grund Anzahl Flüge
- Ausnahme Annullierungen
  - American Eagle Airlines Inc. vergleichsweise viele Annullierungen
  - Delta Air Lines Inc. vergleichsweise wenige Annullierungen



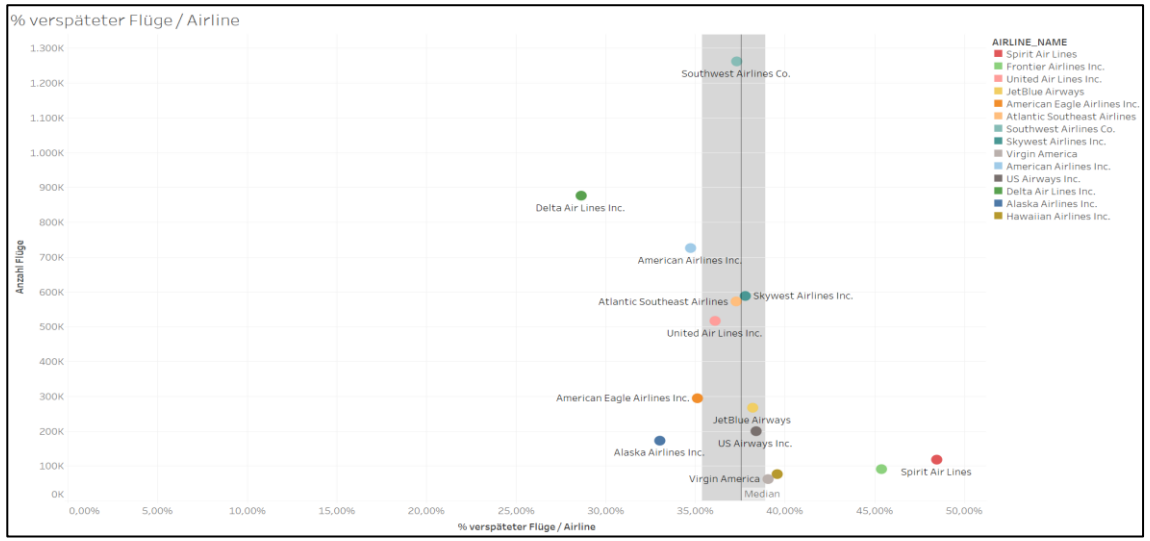


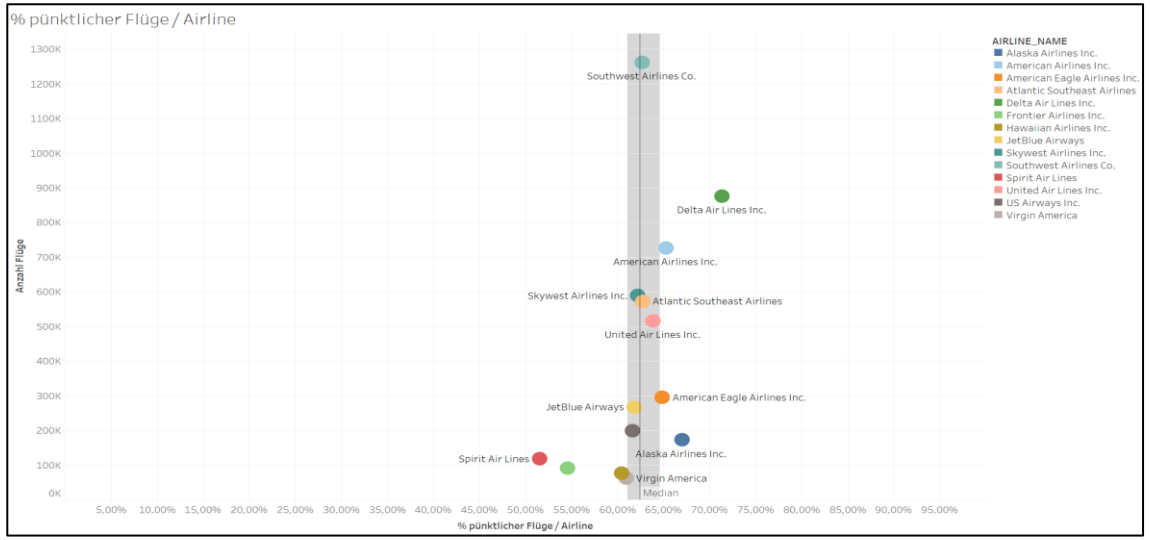




Analyse mit Tableau Desktop Prozentzahlen

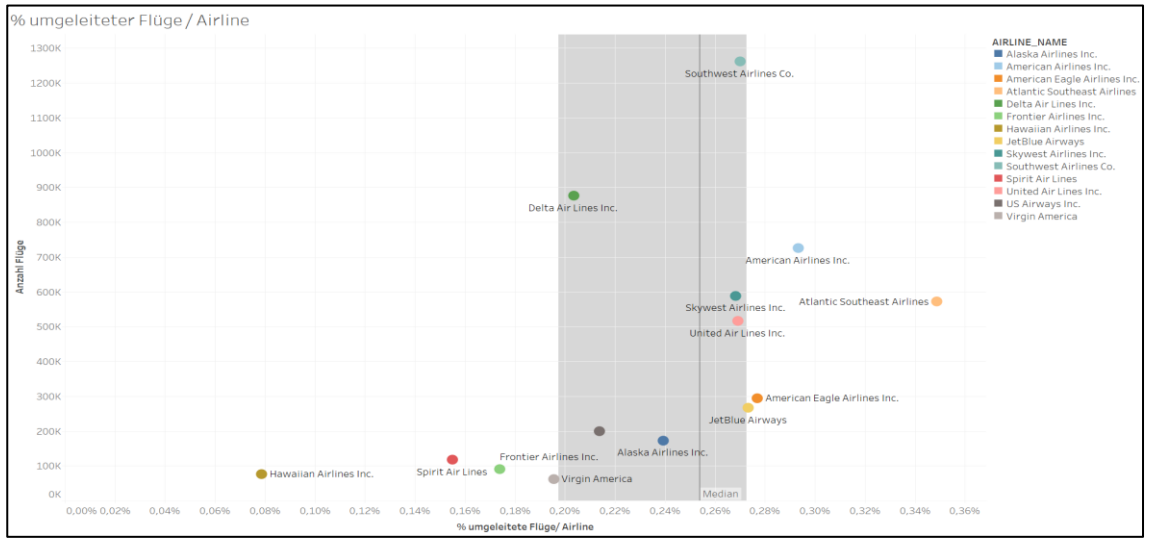
- Hälfte der Airlines zwischen 35% und 38% der Flüge verspätet.
- Am meisten verspätete Airlines
  - Spirit Air Lines, 48% der Flüge
  - Frontier Airlines 45% der Flüge
- Am wenigsten verspätete Airline
  - Delta Air Lines Inc. 29% der Flüge
- Pünktlichste Airlines = verspätete Airlines mit umgekehrten Vorzeichen

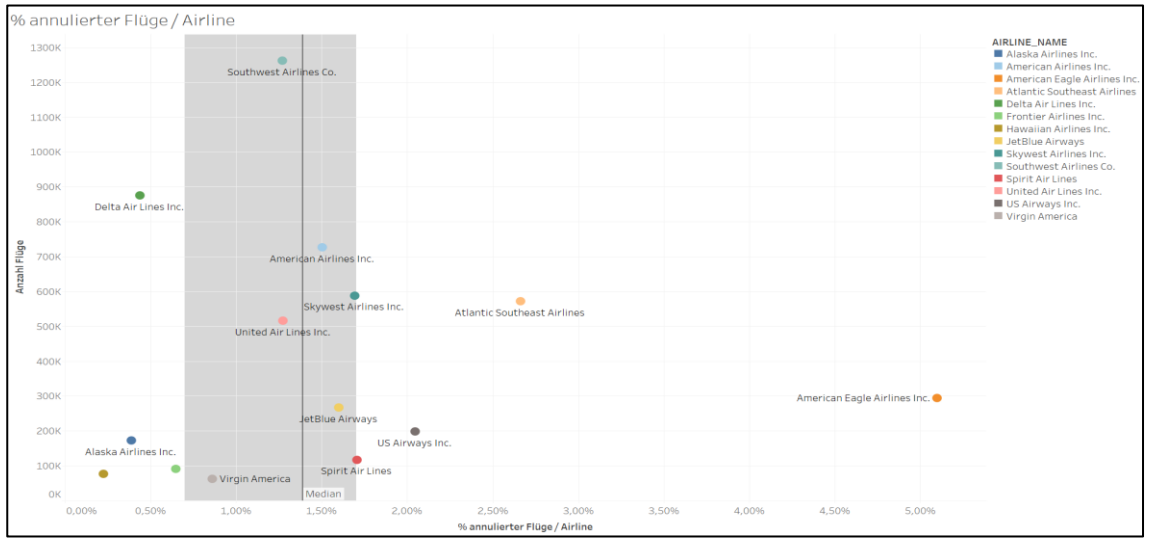




Analyse mit Tableau Desktop Prozentzahlen

- Hälfte der Airlines zwischen 0.2% und 0.27% der Flüge umgeleitet
- Am wenigsten umgeleitete Airline
  - Hawaiian Airlines 0.08% der Flüge
- Am meisten umgeleitete Airline
  - Atlantic Southeast Airlines 0.35% der Flüge
- Hälfte der Airlines zwischen 0.7% und 1.7% der Flüge annulliert
- Am wenigsten annullierte Flüge
  - Hawaiian Airlines 0.22% der Flüge
- Am meisten annullierte Airline
  - American Eagle Airlines 5.1% der Flüge
- Delta Airlines Inc. 875'000 Flüge nur 0.44% annulliert





Zusammenfassung Analyse

- Southwest Airlines Co.
  - Bei absoluten Zahlen immer führend
  - Bei relativer Betrachtung immer in der Hälfte der Airlines
  - Grund Anzahl Flüge
- Spirit Air Lines
  - Am meisten verspätete/am wenigsten pünktliche Flüge
  - Wenig umgeleitete Flüge
  - Im Mittel bei annullierten Flügen
- American Eagle Airlines Inc
  - Relative viele annullierte Fluge
  - Ansonsten im Mittel
- Hawaiian Airlines
  - Wenigsten annullierten und umgeleiteten Flüge
  - Pünktlicher als die Hälfte der Airlines
  - Anzahl der Flüge sehr gering
- Delta Airlines Inc.
  - Am wenigsten verspätete/ am meisten pünktliche Flüge
  - Weniger Annullierungen als die Hälfte der Airlines
  - Im Mittel bei umgeleiteten Flügen
  - Anzahl der Flüge hoch

Fazit / Reflexion

- ETL-Prozess in eigenem Projekt angewendet
- Neue Tools kennengelernt
- Intensivierung von Python
- Anspruchsvoll und zeitintensiv mit unvorhergesehenen Problemen
  - VM-Abgleich, Datenmenge, unvollständige Datensätze, usw.
- Planung von Teamwork und individueller Arbeit
- Neues Wissen erlangt, steigende Lernkurve und **SPASS**!