TECHNICAL EXERCISE

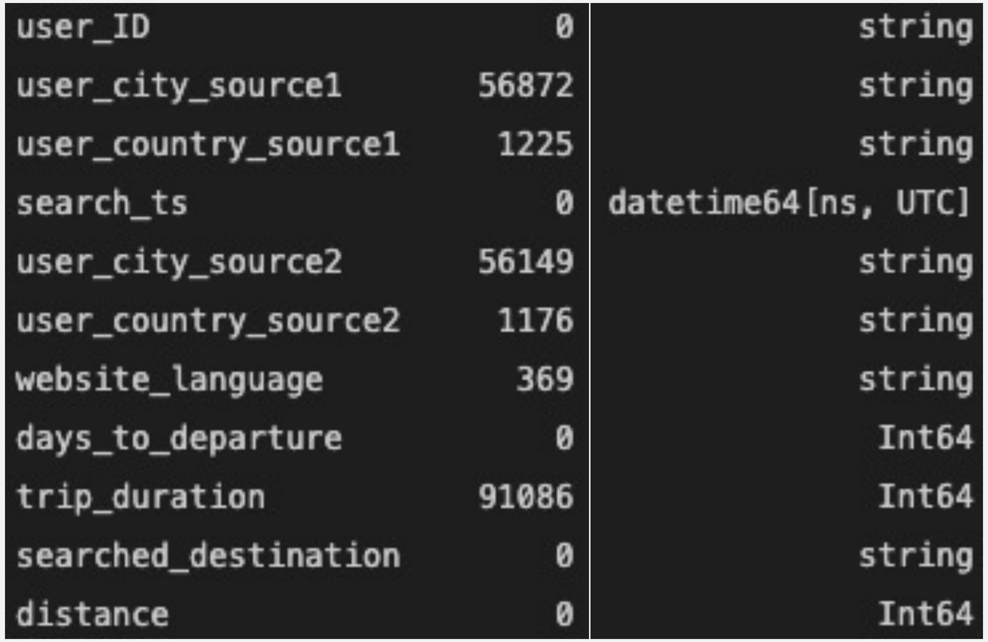
Travel Audience – Senior Data Analyst

Marcos Daniels

DATA QUALITY

OVERVIEW

- Raw data: 231,504 rows; 11 columns
- 138,281 rows containing at least one empty variable
- No duplicates (all columns + all columns except ID)
- “days_to_departure”:
  - 79 rows containing negative values
- “trip_duration”:
  - 0 rows containing negative values
  - 26 rows containing unrealistic values (> 365 days)
- “distance”:
  - 0 rows containing unrealistic values (> 40,000 kms)



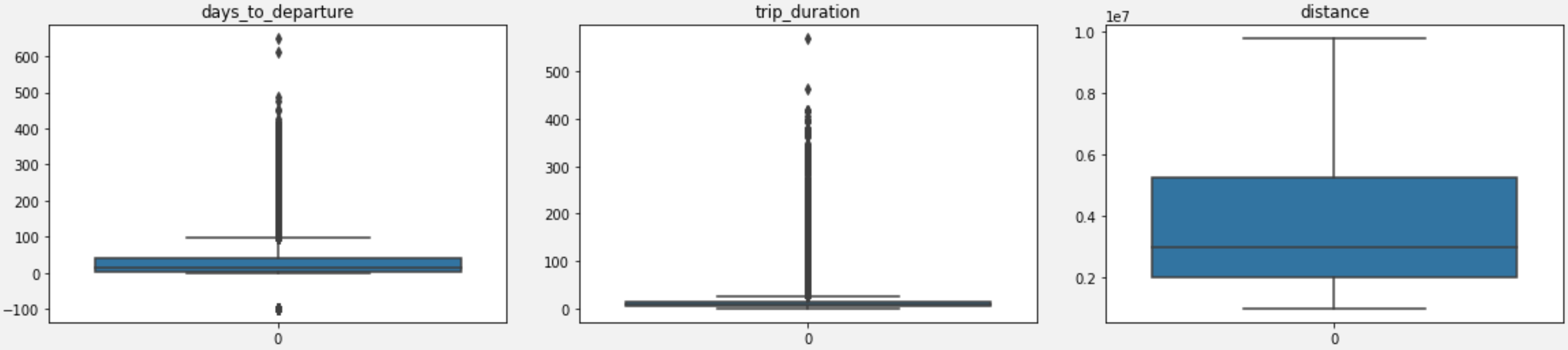
```
user_ID                     0                  string
user_city_source1       56872                  string
user_country_source1     1225                  string
search_ts                   0  datetime64[ns, UTC]
user_city_source2       56149                  string
user_country_source2     1176                  string
website_language          369                  string
days_to_departure           0                   Int64
trip_duration           91086                   Int64
searched_destination        0                  string
distance                    0                   Int64
```

DESCRIPTION (RAW)

	days_to_departure	trip_duration	distance
Unit	days	days	kilometres
Mean	35	14	3681
Std	67	23	2033
Min	-99	0	1007
Max	649	569	9760
75%	41	14	5249

Total rows:
231,504

BOXPLOTS (RAW)

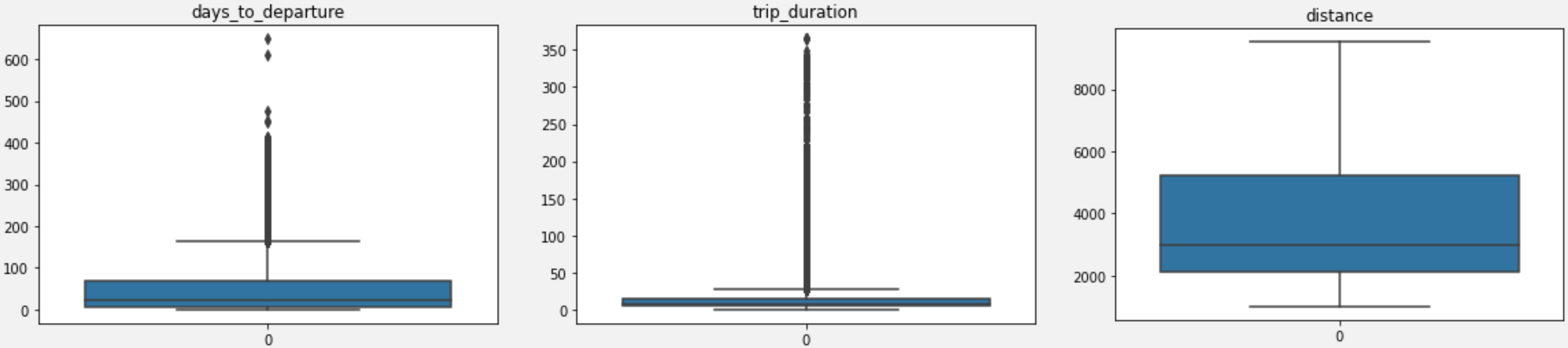


DESCRIPTION (CLEANED)

	days_to_departure	trip_duration	distance
Unit	days	days	kilometres
Mean	53	14	3712
Std	73	22	1939
Min	0	0	1007
Max	649	365	9519
75%	70	14	5222

Total rows:
93,129

BOXPLOTS (CLEANED)



TRENDS

TOTAL

TOP 5 SOURCES

	City 1	Country 1	City 2	Country 2
1.	Dubai	Saudia Arabia	Dubai	Saudia Arabia
2.	Riyadh	United Arab Emirates	Riyadh	United Arab Emirates
3.	Jeddah	Germany	Jeddah	Germany
4.	Dammam	United Kingdom	Dammam	United Kingdom
5.	Abu Dhabi	France	Abu Dhabi	France

TOP 5 SEARCHED DESTINATIONS

1. Istanbul
2. Cairo
3. Abu Dhabi
4. Antalya
5. Hurghada

TOP SOURCES AND DESTINATIONS BY MONTH

	City 1	Country 1	City 2	Country 2	Searched Destination
July	Dubai	Saudi Arabia	Dubai	Saudi Arabia	Cairo
August	Dubai	Saudi Arabia	Dubai	Saudi Arabia	Istanbul
September	Riyadh	Saudi Arabia	Riyadh	Saudi Arabia	Cairo

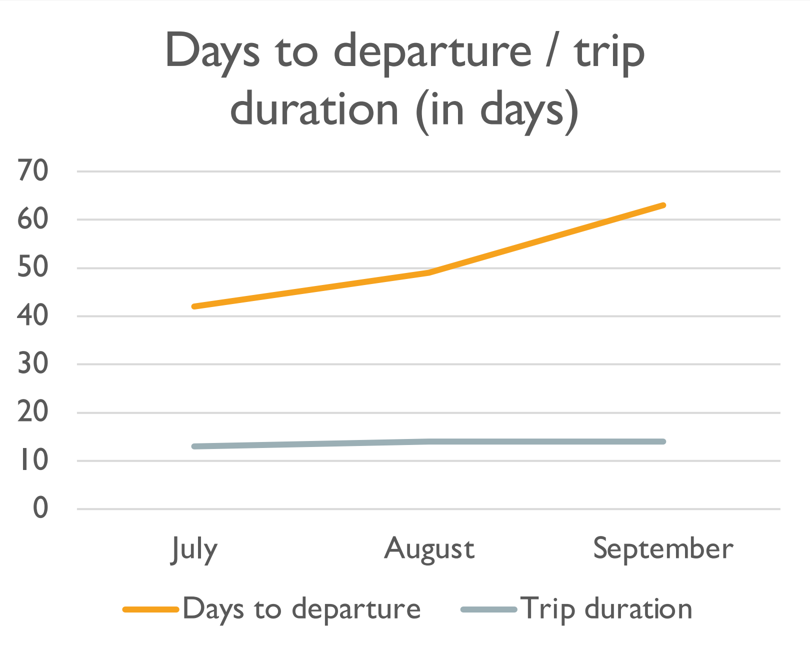
TOP SOURCES AND DESTINATIONS BY MONTH PHASE

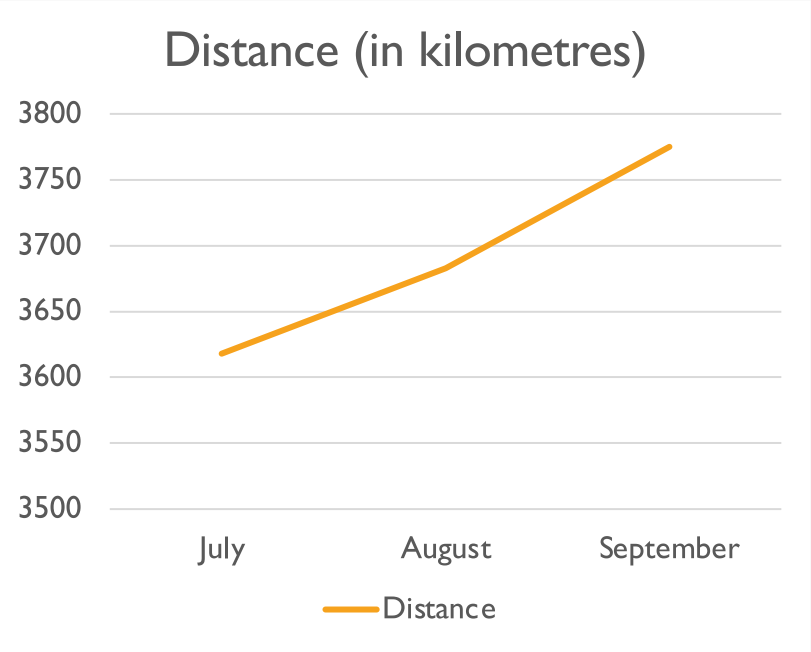
	City 1	Country 1	City 2	Country 2	Searched Destination
Beginning	Riyadh	Saudi Arabia	Dubai	Saudi Arabia	Cairo
Mid	Dubai	Saudi Arabia	Dubai	Saudi Arabia	Istanbul
End	Dubai	Saudi Arabia	Dubai	Saudi Arabia	Istanbul

TOP SOURCES AND DESTINATIONS BY WEEKDAY

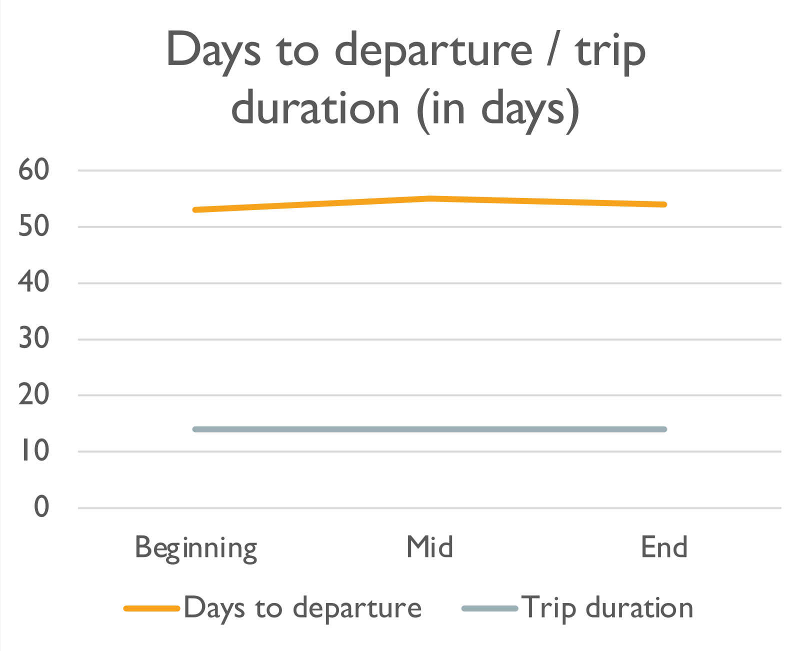
	City 1	Country 1	City 2	Country 2	Searched Destination
Monday	Dubai	Saudi Arabia	Dubai	Saudi Arabia	Istanbul
Tuesday	Riyadh	Saudi Arabia	Riyadh	Saudi Arabia	Istanbul
Wednesday	Dubai	Saudi Arabia	Dubai	Saudi Arabia	Istanbul
Thursday	Dubai	Saudi Arabia	Dubai	Saudi Arabia	Cairo
Friday	Riyadh	Saudi Arabia	Dubai	Saudi Arabia	Istanbul
Saturday	Dubai	Saudi Arabia	Dubai	Saudi Arabia	Cairo
Sunday	Dubai	Saudi Arabia	Dubai	Saudi Arabia	Istanbul

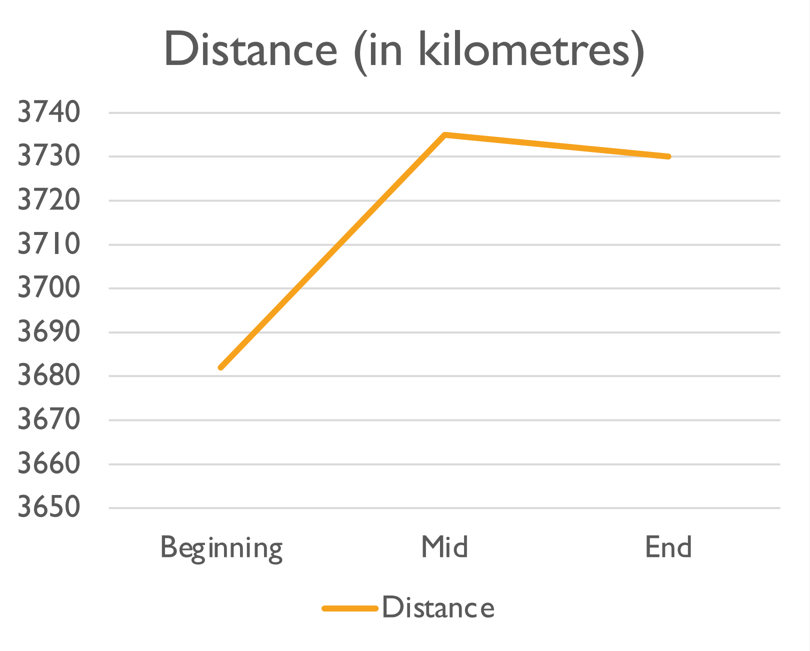
NUMERIC METRICS BY MONTH (AVERAGE)



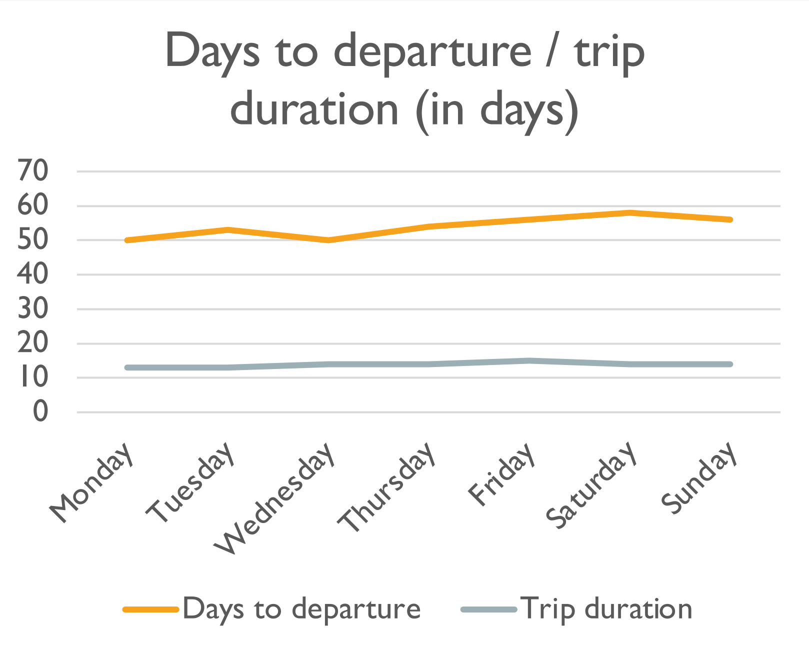


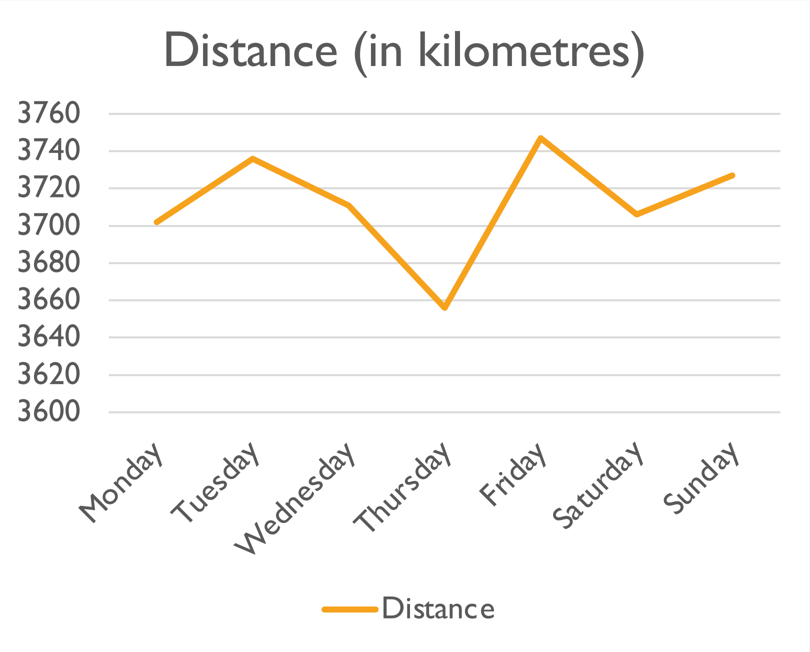
NUMERIC METRICS BY MONTH PHASE (AVERAGE)





NUMERIC METRICS BY WEEKDAY (AVERAGE)





CORRELATIONS

OVERVIEW

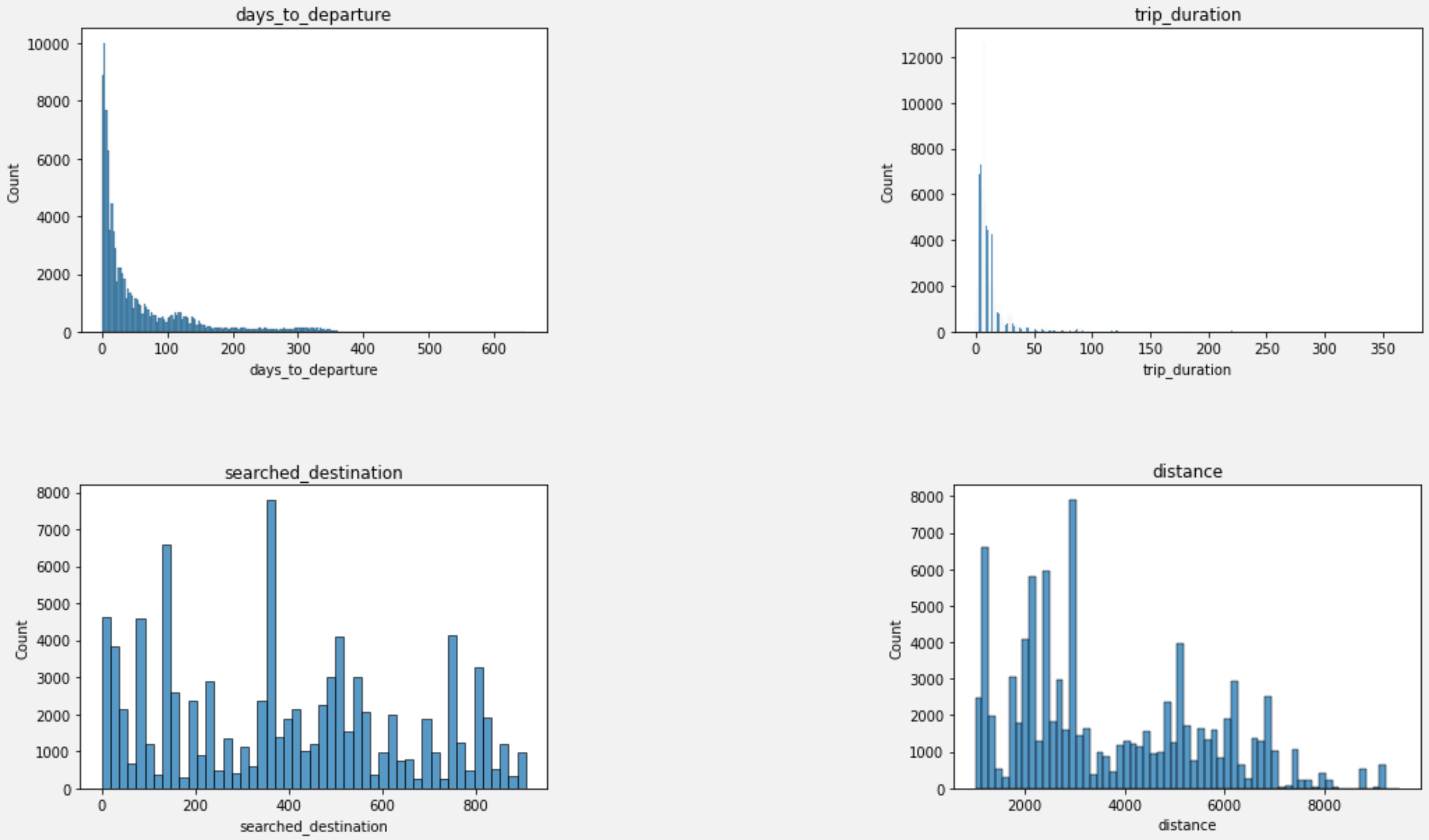
INDEPENDENT VARIABLES

- City 1
- Country 1
- City 2
- Country 2
- Website language
- Month
- Day
- Day (month phase)
- Weekday

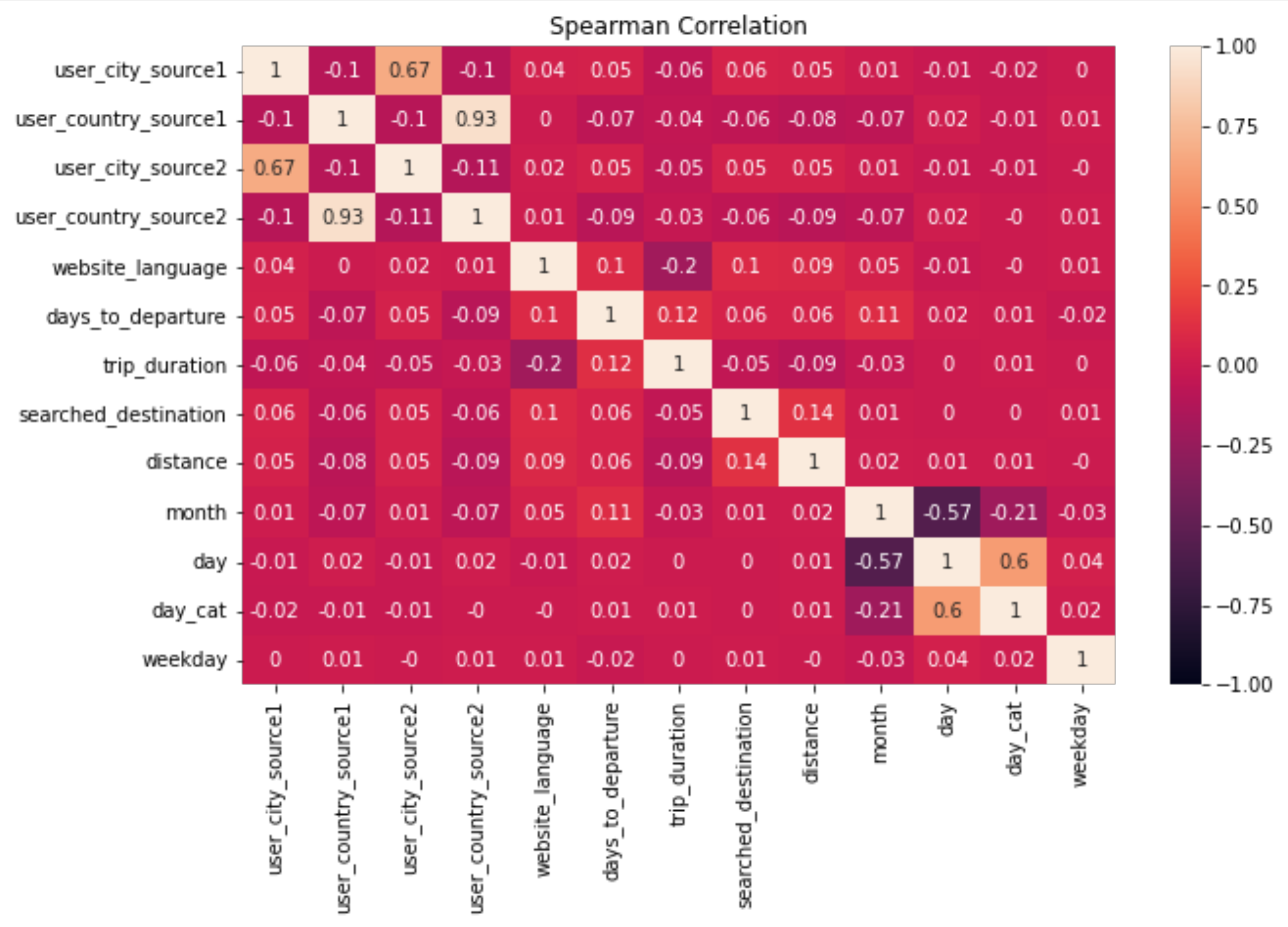
DEPENDENT VARIABLES

- Days to departure
- Trip duration
- Searched destination
- Distance

NORMALLY DISTRIBUTED?



SPEARMAN CORRELATION



Grading Standards	Correlation Degree
$\rho = 0$	no correlation
$0 < \lvert\rho\rvert \leq 0.19$	very week
$0.20 \leq \lvert\rho\rvert \leq 0.39$	weak
$0.40 \leq \lvert\rho\rvert \leq 0.59$	moderate
$0.60 \leq \lvert\rho\rvert \leq 0.79$	strong
$0.80 \leq \lvert\rho\rvert \leq 1.00$	very strong
1.00	monotonic correlation

FURTHER IDEAS

- Fill NA gaps by creating profiles for each user
- Think of an approach to detect anomalies with the help of standard deviation
- Look for other anomalies like multiple changes in city/country/language for the same user within a short time period
- Look for trends by time (hourly / times of day)
- Look for trends by specific cities/countries/languages
- Internalize differences between short-term and long-term trends as well as seasonal adjustments and cycles
- Use other correlation methods like linear regression that also allows for predicting future outcomes

QUESTIONS

- Only data for three months?
- Multiple sources?
- So many NAs?
- No correlations?