ANZ_exploratory_analysis

February 11, 2021

[1]:
```
import pandas as pd
import numpy as np
```

[2]:
```
df = pd.read_excel('ANZ synthesised transaction dataset.xlsx')
```

[3]:
```
df.head(5)
```

[3]:

	status	card_present_flag	bpay_biller_code	account	currency	long_lat	txn_description	merchant_id	merchant_code	first_name	…	age	merchant_suburb	merchant_state	extraction	amount	transaction_id	country	customer_id	merchant_long_lat	movement
0	authorized	1.0	NaN	ACC-1598451071	AUD	153.41 -27.95	POS	81c48296-73be-44a7-befa-d053f48ce7cd	NaN	Diana	…	26	Ashmore	QLD	2018-08-01T01:01:15.000+0000	16.25	a623070bfead4541a6b0fff8a09e706c	Australia	CUS-2487424745	153.38 -27.99	debit
1	authorized	0.0	NaN	ACC-1598451071	AUD	153.41 -27.95	SALES-POS	830a451c-316e-4a6a-bf25-e37caedca49e	NaN	Diana	…	26	Sydney	NSW	2018-08-01T01:13:45.000+0000	14.19	13270a2a902145da9db4c951e04b51b9				
2	authorized	1.0	NaN	ACC-1222300524	AUD	151.23 -33.94	POS	835c231d-8cdf-4e96-859d-e9d571760cf0	NaN	Michael	…	38	Sydney	NSW	2018-08-01T01:26:15.000+0000	6.42	feb79e7ecd7048a5a36ec889d1a94270				
3	authorized	1.0	NaN	ACC-1037050564	AUD	153.10 -27.66	SALES-POS	48514682-c78a-4a88-b0da-2d6302e64673	NaN	Rhonda	…	40	Buderim	QLD	2018-08-01T01:38:45.000+0000	40.90	2698170da3704fd981b15e64a006079e				
4	authorized	1.0	NaN	ACC-1598451071	AUD	153.41 -27.95	SALES-POS	b4e02c10-0852-4273-b8fd-7b3395e32eb0	NaN	Diana	…	26	Mermaid Beach	QLD	2018-08-01T01:51:15.000+0000	3.25	329adf79878c4cf0aeb4188b4691c266				

```
1  Australia  CUS-2487424745     151.21 -33.87    debit
2  Australia  CUS-2142601169     151.21 -33.87    debit
3  Australia  CUS-1614226872     153.05 -26.68    debit
4  Australia  CUS-2487424745     153.44 -28.06    debit

[5 rows x 23 columns]
```

[6]:
```
print(df.shape)
```

```
(12043, 23)
```

[5]:
```
# checking null values
df.isnull().sum()
```

[5]:
```
status                   0
card_present_flag     4326
bpay_biller_code     11158
account                  0
currency                 0
long_lat                 0
txn_description          0
merchant_id           4326
merchant_code        11160
first_name               0
balance                  0
date                     0
gender                   0
age                      0
merchant_suburb       4326
merchant_state        4326
extraction               0
amount                   0
transaction_id           0
country                  0
customer_id              0
merchant_long_lat     4326
movement                 0
dtype: int64
```

The columns with missing values are card_present_flag, bpay_biller_code, merchant_id, merchant_code, merchant_suburb, merchant_state and merchant_long_lat.

[13]:
```
# checking the types of transcations where merchant_id is null
df[df.merchant_id.isnull().values]['txn_description'].unique()
```

[13]:
```
array(['PAYMENT', 'INTER BANK', 'PAY/SALARY', 'PHONE BANK'], dtype=object)
```

[14]:
```
# checking the types of transcations where merchant_id is not null
df[df.merchant_id.notnull().values]['txn_description'].unique()
```

[14]:
```
array(['POS', 'SALES-POS'], dtype=object)
```

Merchant_id is only missing for non-merchant involved transactions, such as Payment, inter bank, etc.

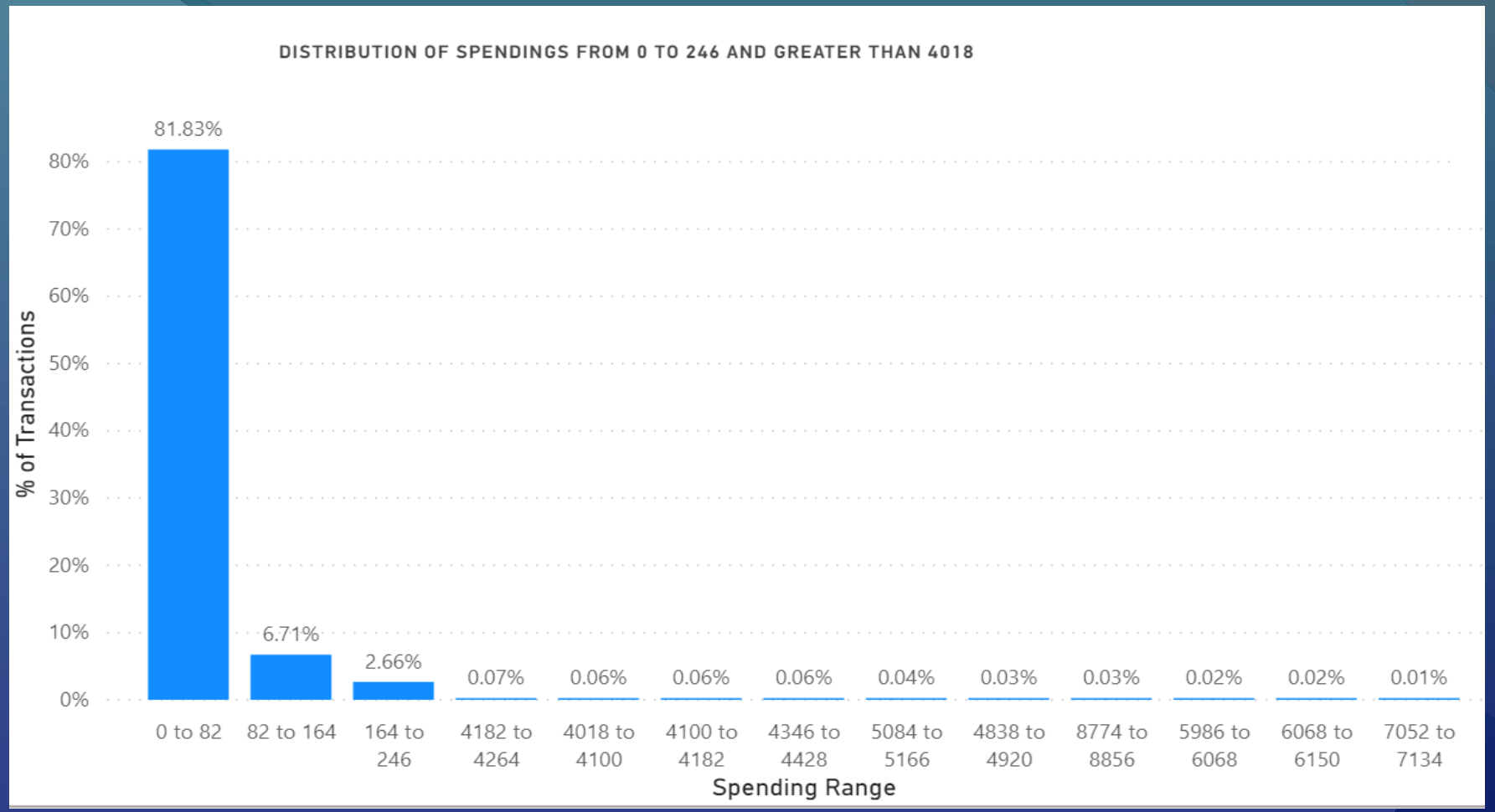
Columns related to merchants such as merchant_id, merchant_suburb, merchant_state, merchant_long_lat all have 4326 missing values, and they are missing only for non merchant involved transactions. Also, card_present_flag has 4326 missing values. So, this column has also some kind of relationship with the presence of merchants.

From the initial analysis, it can be considered that the bpay_biller_code and merchant_code which have too many missing values, 11158 and 11160 respectively, are irrelevant and can be dropped or imputated in the next stage of our analysis process.

[ ]:
```

```

A N Z ' s T r a n s a c t i o n D a t a



INSIGHTS

91.2%

Transactions in the range 0 to 246 occurred 11079 times which is 91.2% of total transactions.

0.39%

Transactions greater than 4018 occurred 47 times which is 0.39% of total transactions.

A N Z ' s T r a n s a c t i o n D a t a

txn_description	amount	Median Avg transaction by type	Count of Id
INTER BANK	64,331.00	39.00	742
PAY/SALARY	1,676,576.85	1,626.48	883
PAYMENT	201,794.00	42.50	2600
PHONE BANK	10,716.00	43.00	101
POS	152,861.24	19.43	3783
SALES-POS	157,005.11	20.04	3934
Total	**2,263,284.20**	**29.00**	**12043**

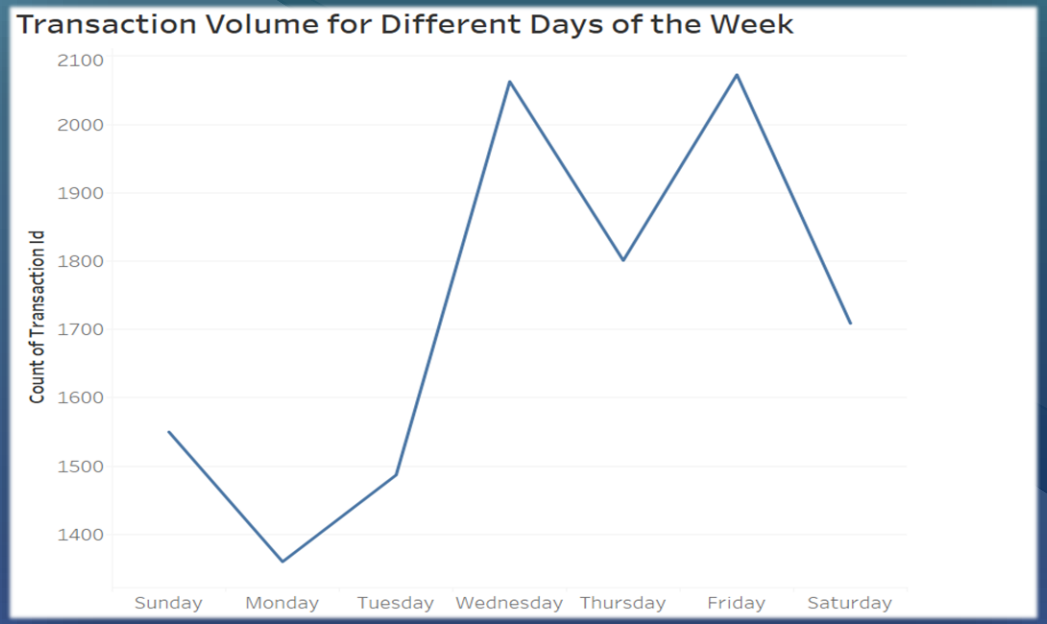
19.43

The median average transaction amount for POS is 19.43 AUD which is less compared to all transaction types.

43

Despite having the lowest number of transactions (101), phone banking has the second-highest average median transaction amount (43 AUD) only less than salary transaction, which is 1623.48 AUD.

A N Z ' s T r a n s a c t i o n D a t a



Transaction volumes start to rise after Monday till Friday except for Thursday before falling on Saturday and Sunday.

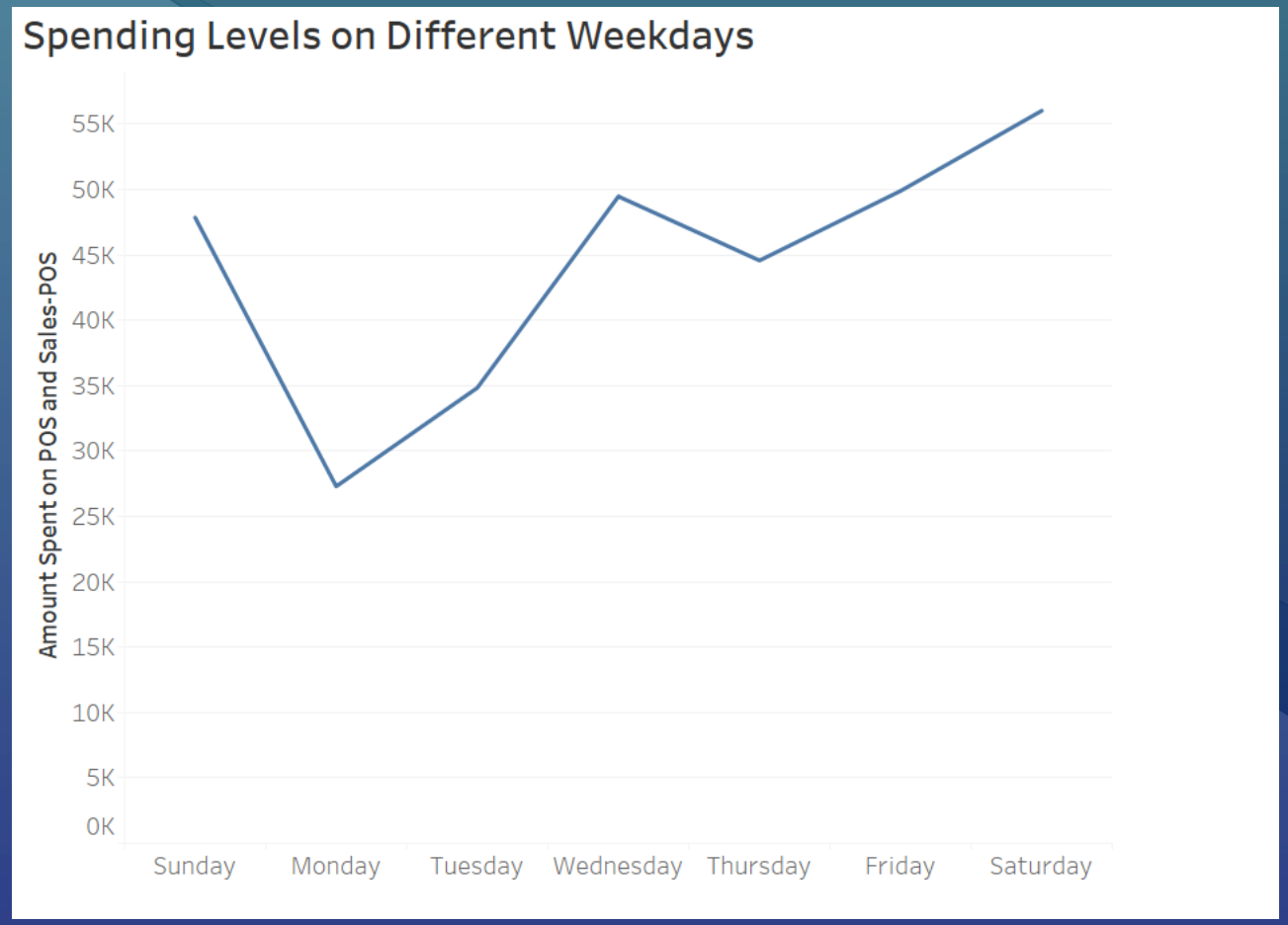
For both peak days (Wednesday and Friday), transaction volumes for every transaction type were higher compared to other days.

Although Saturday and Sunday saw a dip in transaction volumes, POS and Sales-POS transactions remained at their highest level.

Categories of Transactions on Different Weekdays

Txn Description	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
INTER BANK	122	99	123	169	87	80	62
PAY/SALARY		207	160	172	143	201	
PAYMENT	297	311	345	441	442	470	294
PHONE BANK	5	1	18	30	26	5	16
POS	533	354	414	602	573	655	652
SALES-POS	593	388	427	649	530	662	685

A N Z ' s T r a n s a c t i o n D a t a



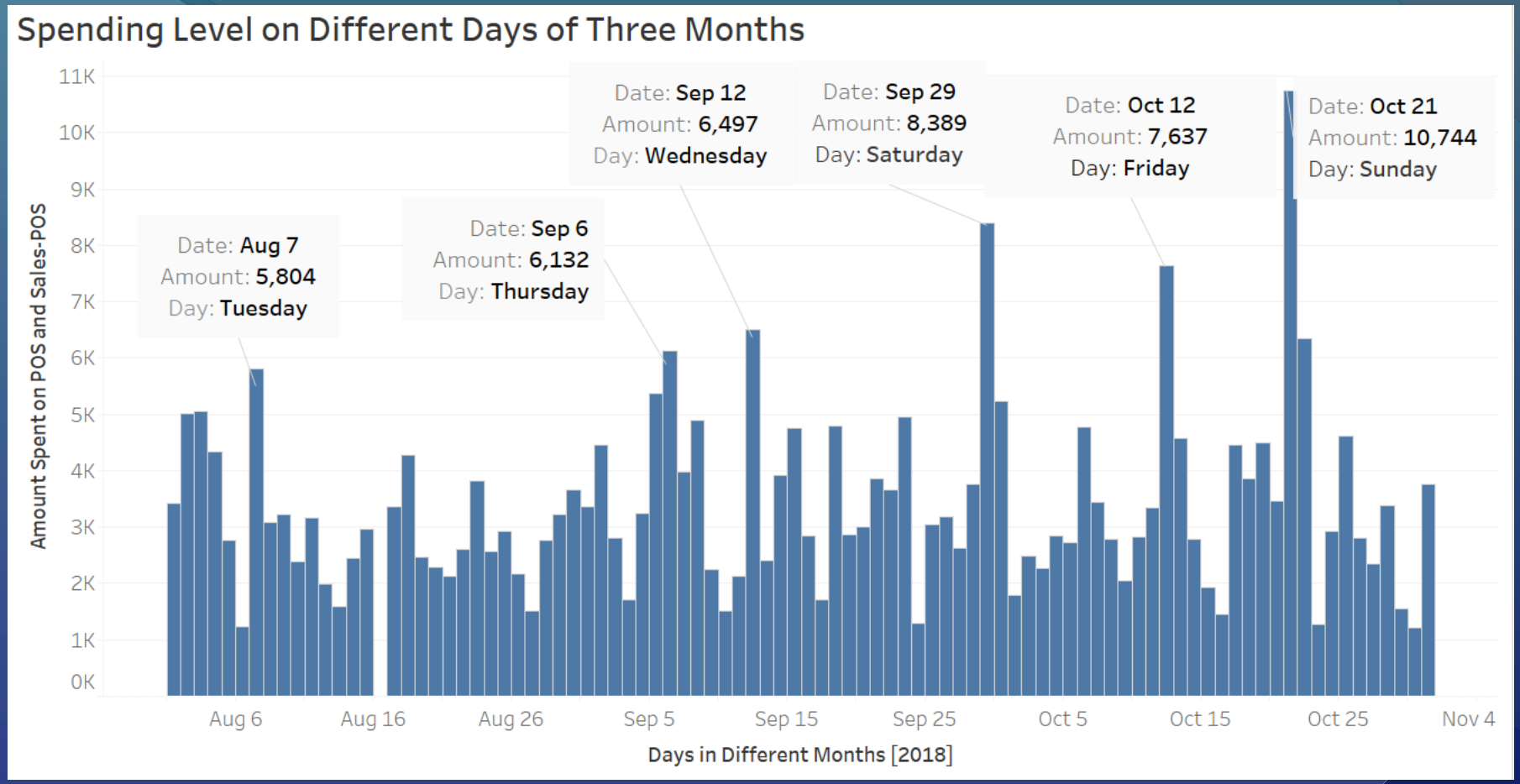
INSIGHTS

Trend

Overall, the spending follows an upward trend from Monday to Saturday except for Thursday before shrinking on Sunday.

However, it needs to be checked if outliers have any effect on this trend or not.

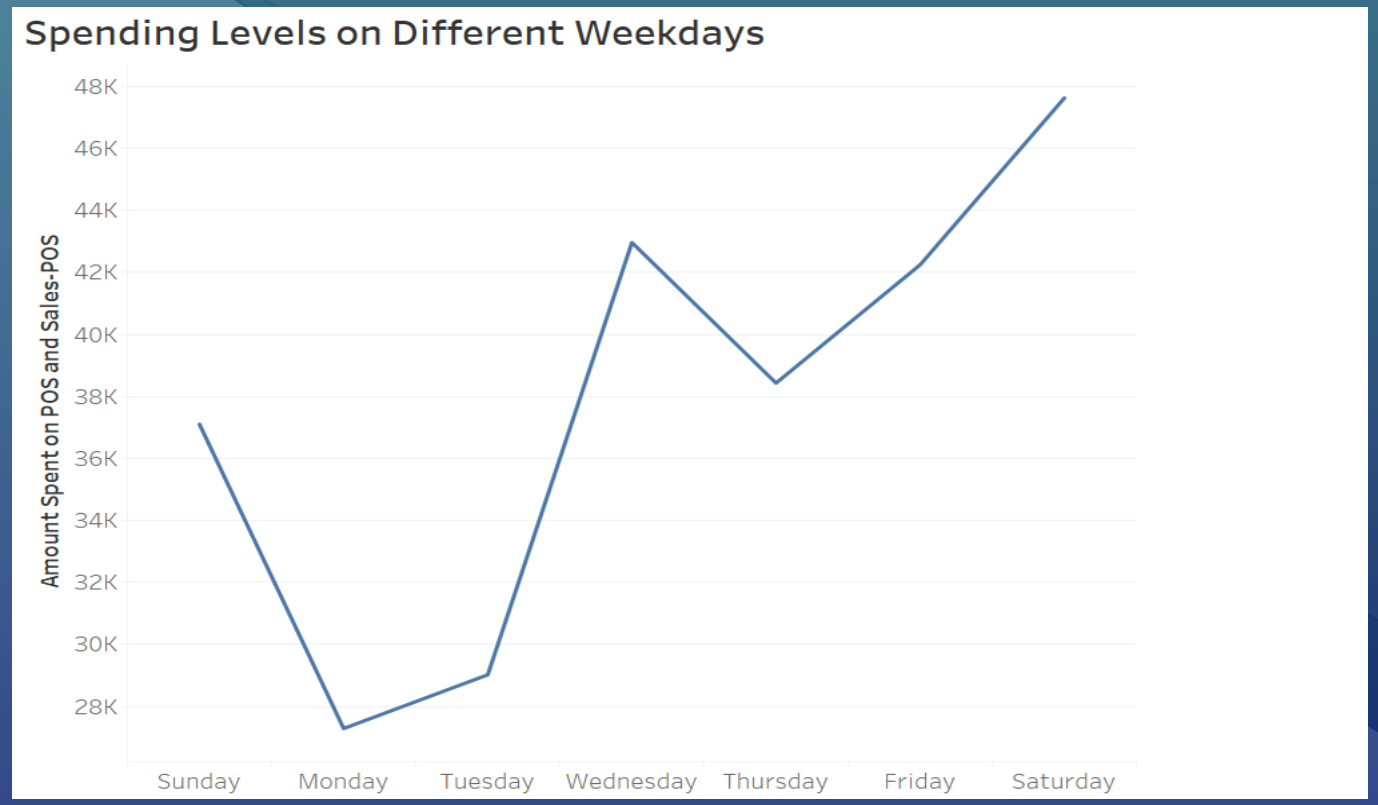
A N Z ' s T r a n s a c t i o n D a t a



Outliers

From the bar graph on the left, it is observed that there are outlier records on every day of the week except for Monday. These outliers might have biased the trend towards a certain direction.

A N Z ' s T r a n s a c t i o n D a t a



Same

The trend stayed the same even after the removal of outlier records.

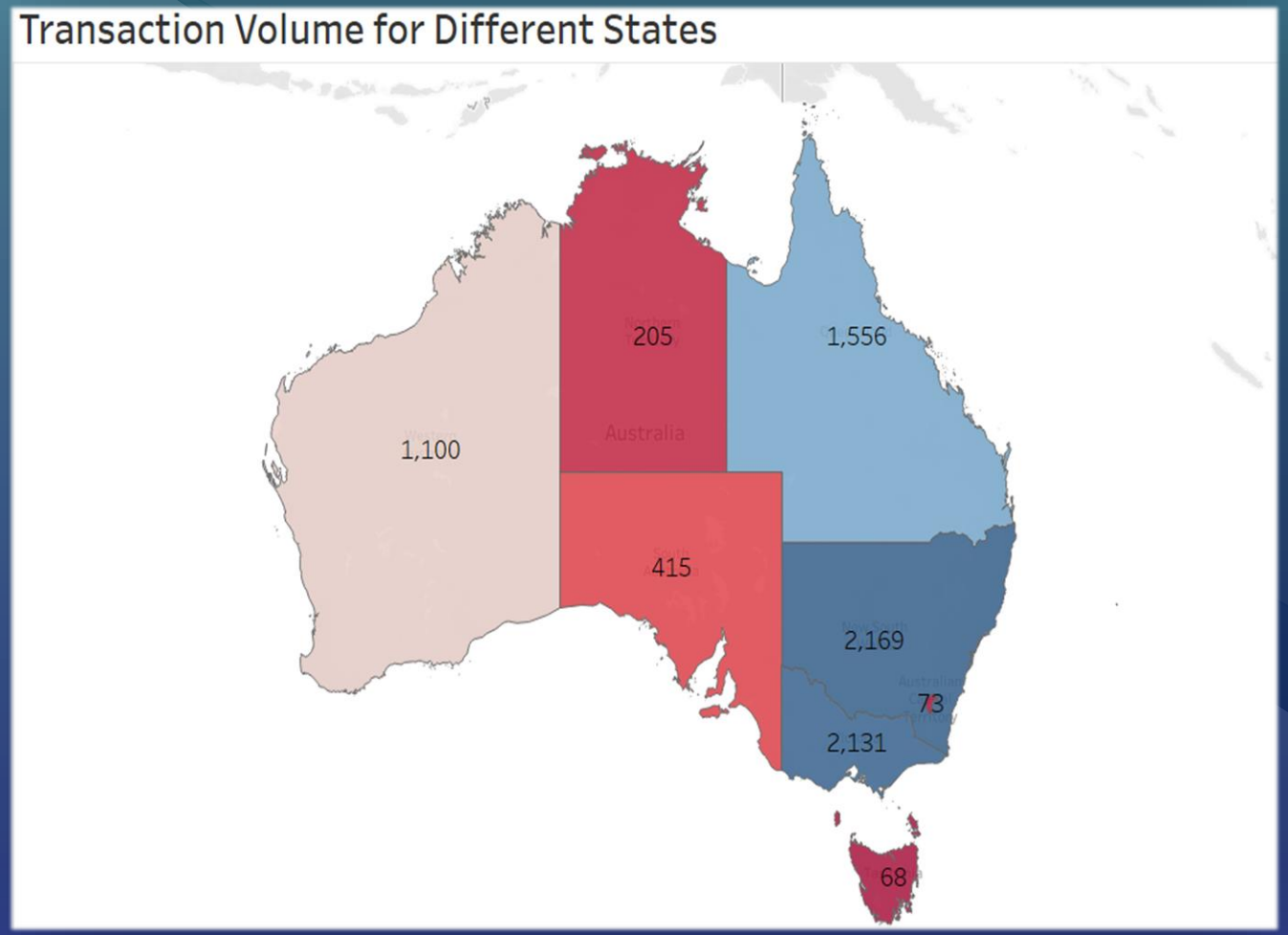
Saturday

Saturday is when spending level is maximum by people at a total of 47,621 AUD.

Low Trading Days

The spending level on Monday and Tuesday showed the two lowest values which are below 30,000 AUD.

ANZ's Transaction Data



High

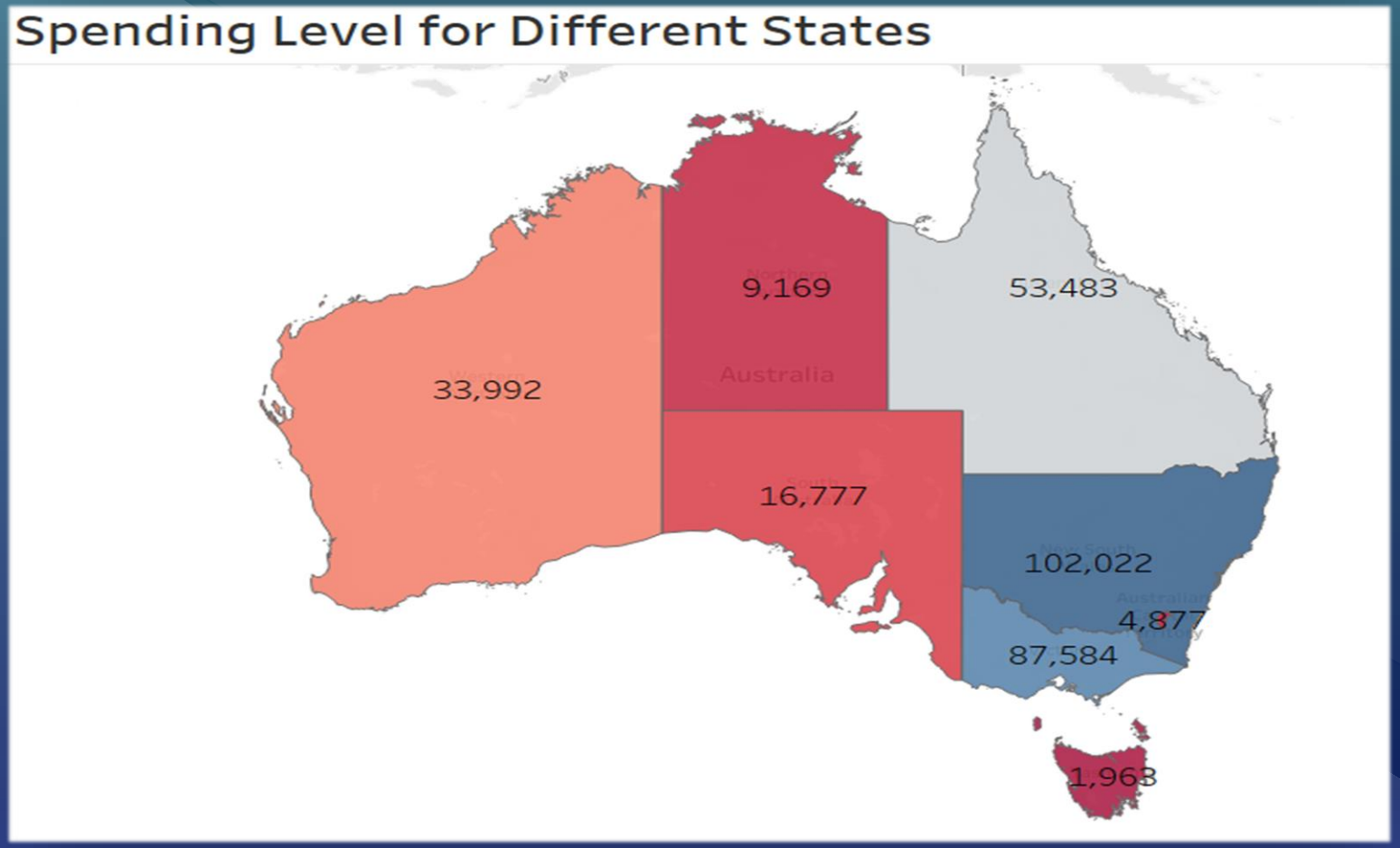
The three most populous states (NSW, Victoria and Queensland) in Australia witnessed higher transaction volumes with NSW topping the list at 2,169 followed by Victoria (2,131) and Queensland (1,556).

Low

ACT and Tasmania had two lowest transaction volumes at 78 and 68, respectively.

INSIGHTS

ANZ's Transaction Data



High

Spending is highest in NSW (102,022 AUD). Victoria and Queensland came second and third with 87,584 AUD and 53,483 AUD respectively.

Low

The spending level for ACT and Tasmania are lowest at 4,877 AUD and 1,968 AUD respectively.