Data Analysis

September 7, 2020

[3]:
```
import pandas as pd
import os
```

[5]:
```
file_name = os.listdir()[0]
print(file_name)
```

```
2019 Winter Data Science Intern Challenge Data Set - Sheet1.csv
```

[7]:
```
df = pd.read_csv(file_name)
df.head()
```

[7]:

	order_id	shop_id	user_id	order_amount	total_items	payment_method	created_at
0	1	53	746	224	2	cash	2017-03-13 12:36:56
1	2	92	925	90	1	cash	2017-03-03 17:38:52
2	3	44	861	144	1	cash	2017-03-14 4:23:56
3	4	18	935	156	1	credit_card	2017-03-26 12:43:37
4	5	18	883	156	1	credit_card	2017-03-01 4:35:11

[20]:
```
"""
Naive Average
"""
print("AOV: ",df["order_amount"].mean())
```

```
AOV:  3145.128
```

As we can see that if we do a simple average operation it returns the miscalculate values form the original document. This is due to the fact that calulation does not account for the total items in order. We can avoid this by dividing the order amount by total items then calculating the avarage from that.

[21]:
```
average_price = df["order_amount"]/df["total_items"]
print("Actual AOV" , average_price.mean())
```

```
Actual AOV 387.7428
```

The actual AOV comes out to 387$ much more realistic than the the stated value.

[ ]: