Nyambok Julius Nyerere

IBM Challenge Documentary Walkthrough

1.Find the top 7 most travelled routes for a Sunday on average, indicate the average of each and rank them in decreasing order.

Before starting number 1, when working with dates you have to ensure that all the date formats are uniform because when doing an overview of the data, I noticed that some date formats are inconsistent i.e., dates like 11-10-17 which represent 10th November 2017 in the mm/dd/yyyy and also represent 11th October 2017 in the dd/mm/yyyy.

I used the `pd.to_datetime()` to ensure that date formats are consistent so that the results are not skewed in anyway. I used `dt.day_name()` to change the dates to days of the week.

```
data['day'] = data['travel_date'].dt.day_name()
data.head()
```

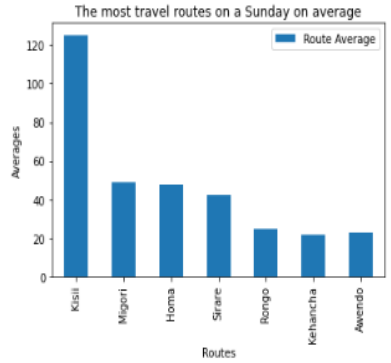
	ride_id	seat_number	payment_method	payment_receipt	travel_date	travel_time	travel_from	travel_to	car_type	max_capacity	day
0	1442	15A	Mpesa	UZUEHCBUSO	2017-10-17	7:15	Migori	Nairobi	Bus	49	Tuesday
1	5437	14A	Mpesa	TIHLBUSGTE	2017-11-19	7:12	Migori	Nairobi	Bus	49	Sunday
2	5710	8B	Mpesa	EQX8Q5G19O	2017-11-26	7:05	Keroka	Nairobi	Bus	49	Sunday
3	5777	19A	Mpesa	SGP18CL0ME	2017-11-27	7:10	Homa Bay	Nairobi	Bus	49	Monday
4	5778	11A	Mpesa	BM97HFRGL9	2017-11-27	7:12	Migori	Nairobi	Bus	49	Monday

I then filtered out the travelers who travelled on a Sunday. On close observation, I noticed that directly dividing the total number of rows (travelers of every route who travelled on a Sunday) by the number of Sundays to find the average by route would be inaccurate because the average would not be properly represented because of 'double counting'(travelers travelling on the same Sunday would not be properly represented).

I instead opted to filter out the travelers by route name, find out on how many distinct Sundays did travelers use the route, divide it by the number of travelers who used the route on all Sundays to find the average. i.e., The average number of travelers travelling Kisii **on any given Sunday** is 125.27 (Rounded down to 125 since there can't be a '0.27th' of a human being)

```
route_averages.iloc[0:7].plot(kind='bar',x='Route Name',y='Route Average',xlabel='Routes',ylabel='Averages',title='The most travel routes on a Sunday on average')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f63784bbf50>
```



2.What is the probability that a passenger travelling from Kijauri will take a Shuttle if they depart before 07:30?

I filtered out travelers that had plied the Kijauri route using a shuttle and traveled before 7.30 and found the probability of anyone taking the shuttle to Kijauri before 7.30 against the total number of travelers taking the route by shuttle to Kijauri at any other time

```
[13] kijauri_route_before_7_30=kijauri_route.query('travel_time<"7:30" and car_type=="shuttle"')
     len(kijauri_route_before_7_30)
```

```
487
```

```
[14] probability=((len(kijauri_route_before_7_30))/(len(kijauri_route)))
     probability
```

```
0.472356935014549
```

3. ***The Sequence 'MK' appears in a payment reference. Based on the distribution of characters in all the payment references what do you think is the most probable next character (if any)?***

I looped through the 'payment receipt' column to find any receipt with a sequence of 'MK' at any point of the receipt. MK can occur at the end, (takes up the last two characters at the end) meaning if this happens frequently, the most probable character is 'null'

Therefore, MK occurs at the end of the character sequence more than any position meaning no character succeeds it.

```
characters=[]

for a in receipt_with_mk:
    position = a.find('MK')
    if position<8:
        characters.append(a[position+2])
    else:
        characters.append('No character proceeds here')
```

```
[22] chars_after_MK = pd.DataFrame(characters,columns=['count'])
```

```
chars_after_MK['count'].value_counts()
```

```
No character proceeds here    52
L                             20
Y                             20
H                             19
R                             19
P                             18
M                             18
```