In [2]:

```
import pandas as pd
import seaborn as sns
```

In [3]:

```
train = pd.read_csv("train2.csv") # loading the train1 dataset from step_2
```

In [4]:

```
train.head()
```

Out[4]:

	order_no	user_id	vehicle_type	platform_type	customer_type	placed_day	pl
0	Order_No_4211	User_Id_633	Bike	3	Business	9	
1	Order_No_25375	User_Id_2285	Bike	3	Personal	12	
2	Order_No_1899	User_Id_265	Bike	3	Business	30	
3	Order_No_9336	User_Id_1402	Bike	3	Business	15	
4	Order_No_27883	User_Id_1737	Bike	1	Personal	13	

5 rows × 32 columns

In [5]:

```
train.columns
```

Out[5]:

```
Index(['order_no', 'user_id', 'vehicle_type', 'platform_type', 'customer_type'
,
       'placed_day', 'placed_wkday', 'placed_time', 'confirmed_day',
       'confirmed_wkday', 'confirmed_time', 'arrive_pickup_day',
       'arrive_pickup_wkday', 'arrive_pickup_time', 'pickup_day',
       'pickup_wkday', 'pickup_time', 'delivered_day', 'delivered_wkday',
       'delivered_time', 'distance_km', 'temp', 'pickup_lat', 'pickup_long',
       'delivered_lat', 'delivered_long', 'Rider Id', 'time_pickup_to_arrival'
,
       'No_Of_Orders', 'Age', 'Average_Rating', 'No_of_Ratings'],
      dtype='object')
```

In [6]:

```
train['confirmed_wkday'].value_counts() #seeing how many orders are placed wi
```

Out[6]:

```
4    4229
5    3993
2    3959
3    3823
1    3788
6    1223
7     186
Name: confirmed_wkday, dtype: int64
```

From here there are:

- 3788 orders on Monday
- 3959 orders on Tuesday
- 3823 orders on Wednesday
- 4229 orders on Thursday
- 3993 orders on Friday
- 1223 orders on Saturday
- 186 orders on Sunday

In [19]:

```
train['placed_day'].value_counts() #seeing how many orders are placed within
```

Out[19]:

```
8      848
7      822
13     811
14     804
6      794
28     784
4      769
18     769
15     762
11     752
5      747
3      718
30     714
10     709
25     691
29     685
27     670
9      667
12     666
22     650
21     649
20     643
26     639
2      602
17     593
24     591
19     589
16     565
23     563
1      482
31     453
Name: placed_day, dtype: int64
```

Categorical Variables

In [31]:

```
pd.crosstab(train.vehicle_type, train.customer_type) #Let's begin by taking a
```

Out[31]:

customer_type	Business	Personal
vehicle_type		
Bike	17384	3817

In [43]:

```
df = pd.read_csv("train2.csv")
```

In [45]:

```
df
```

Out[45]:

	order_no	user_id	vehicle_type	platform_type	customer_type	placed_da
0	Order_No_4211	User_Id_633	Bike	3	Business	9
1	Order_No_25375	User_Id_2285	Bike	3	Personal	1
2	Order_No_1899	User_Id_265	Bike	3	Business	3
3	Order_No_9336	User_Id_1402	Bike	3	Business	1
4	Order_No_27883	User_Id_1737	Bike	1	Personal	1
...	...	...	...	...	...	..
21196	Order_No_8834	User_Id_2001	Bike	3	Personal	2
21197	Order_No_22892	User_Id_1796	Bike	3	Business	1
21198	Order_No_2831	User_Id_2956	Bike	3	Business	
21199	Order_No_6174	User_Id_2524	Bike	1	Personal	4
21200	Order_No_9836	User_Id_718	Bike	3	Business	2

21201 rows × 32 columns

In [48]:

```
df.groupby('delivered_wkday').Age.median() #We can also look at how a numeric
```

Out[48]:

```
delivered_wkday
1    872.0
2    872.0
3    874.0
4    872.0
5    846.0
6    824.0
7    900.0
Name: Age, dtype: float64
```

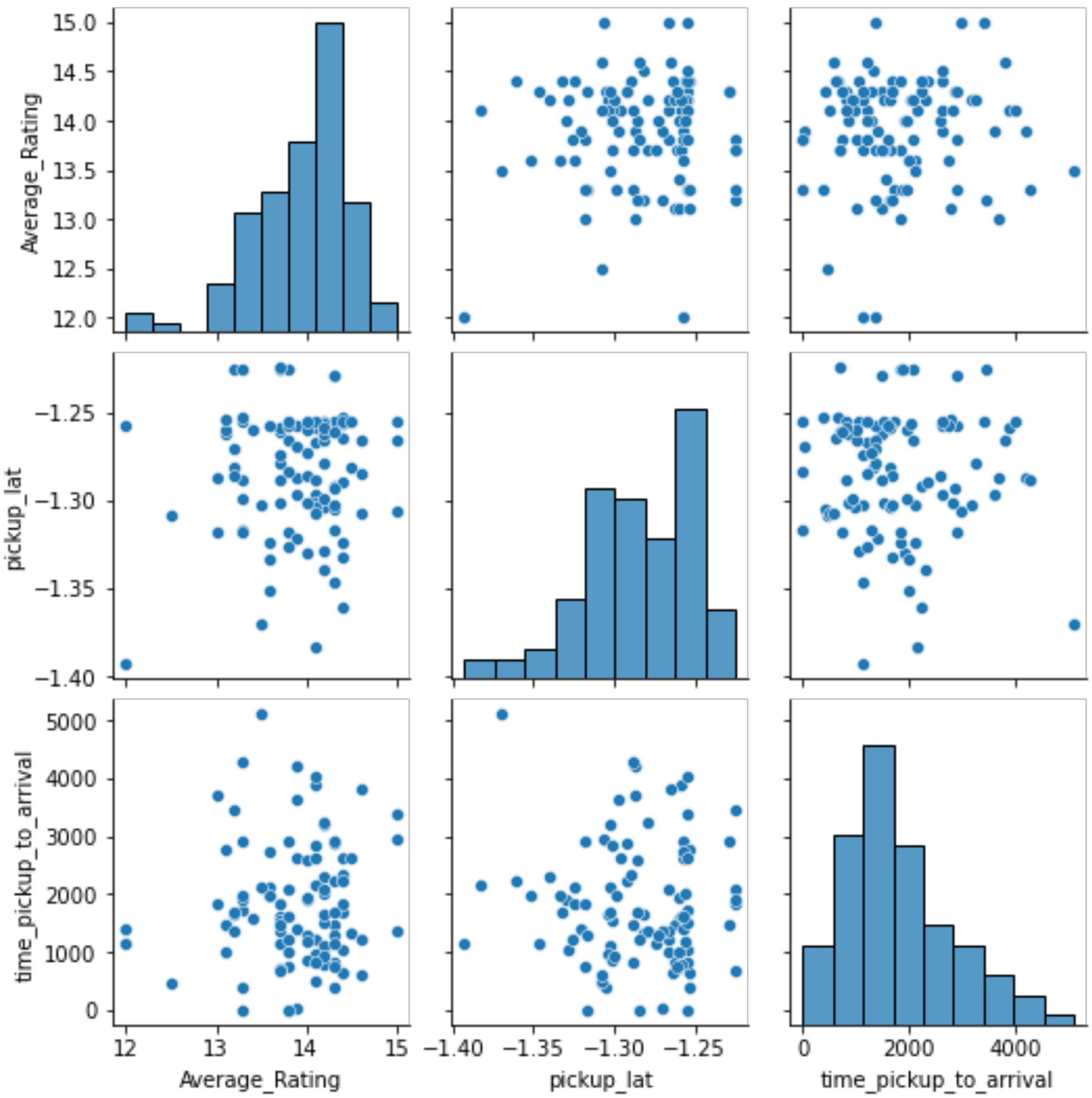
Numerical Variables

In [54]:

```
sns.pairplot(df[['Average_Rating', 'pickup_lat', 'time_pickup_to_arrival']].h
```

Out[54]:

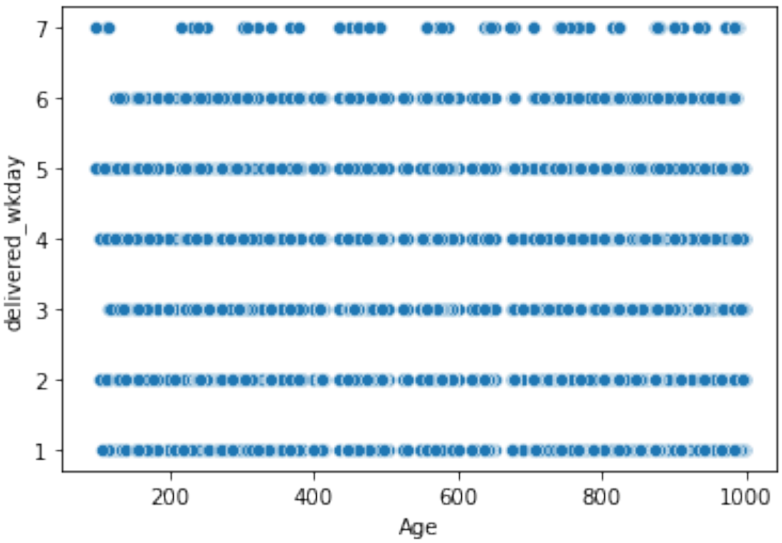
```
<seaborn.axisgrid.PairGrid at 0x7ffc782e5f40>
```



In [57]:

```
sns.scatterplot(x="Age", y="delivered_wkday", data=df[df.Age < 1000])
```

```
Out[57]: <AxesSubplot:xlabel='Age', ylabel='delivered_wkday'>
```



```
In [61]: Age = pd.read_csv("train2.csv")
```

```
In [62]: Age
```

Out[62]:

	order_no	user_id	vehicle_type	platform_type	customer_type	placed_da
0	Order_No_4211	User_Id_633	Bike	3	Business	9
1	Order_No_25375	User_Id_2285	Bike	3	Personal	1
2	Order_No_1899	User_Id_265	Bike	3	Business	3
3	Order_No_9336	User_Id_1402	Bike	3	Business	1
4	Order_No_27883	User_Id_1737	Bike	1	Personal	1
...	...	...	...	...	...	..
21196	Order_No_8834	User_Id_2001	Bike	3	Personal	2
21197	Order_No_22892	User_Id_1796	Bike	3	Business	1
21198	Order_No_2831	User_Id_2956	Bike	3	Business	
21199	Order_No_6174	User_Id_2524	Bike	1	Personal	4
21200	Order_No_9836	User_Id_718	Bike	3	Business	2

21201 rows × 32 columns

In [64]:

```
correlations = Age[Age_cols].corr()
correlations
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
/var/folders/1b/x8dgds9541d_h48rwxtt2tkh0000gn/T/ipykernel_10759/568613662.py
in <module>
----> 1 correlations = Age[Age_cols].corr()
      2 correlations

NameError: name 'Age_cols' is not defined
```

In [ ]: