

Assignment Report for Data Analytics Intern

Name: Abhiram Shanmuga

abhiramchowdhary9@gmail.com

+91 8197790005

Introduction: This study describes how to find popular meals, investigate demographic factors impacting user behaviour, and analyse the link between cooking sessions and user orders. A dataset containing details on user orders, cooking sessions, and demographics is used for the investigation.

1. Data Preprocessing:

Step 1:

We put three Excel files with user information, cooking session information, and order information into several Pandas Data Frames to start our investigation. We were able to get a basic grasp of each dataset's structure, column names, and data types by looking at the first few rows. This phase was essential for organising the next stages of our study and for seeing any possible problems with the quality of the data.

	User ID	User Name	Age	Location	Registration Date	Phone	
0	U001	Alice Johnson	28	New York	2023-01-15	123-456-7890	
1	U002	Bob Smith	35	Los Angeles	2023-02-20	987-654-3210	
2	U003	Charlie Lee	42	Chicago	2023-03-10	555-123-4567	
3	U004	David Brown	27	San Francisco	2023-04-05	444-333-2222	
4	U005	Emma White	30	Seattle	2023-05-22	777-888-9999	

	Email	Favorite Meal	Total Orders
0	alice@email.com	Dinner	12
1	bob@email.com	Lunch	8
2	charlie@email.com	Breakfast	15
3	david@email.com	Dinner	10
4	emma@email.com	Lunch	9

	Session ID	User ID	Dish Name	Meal Type	Session Start	
0	S001	U001	Spaghetti	Dinner	2024-12-01 19:00:00	
1	S002	U002	Caesar Salad	Lunch	2024-12-01 12:00:00	
2	S003	U003	Grilled Chicken	Dinner	2024-12-02 19:30:00	
3	S004	U001	Pancakes	Breakfast	2024-12-02 07:30:00	
4	S005	U004	Caesar Salad	Lunch	2024-12-03 13:00:00	

	Session End	Duration (mins)	Session Rating
0	2024-12-01 19:30:00	30	4.5
1	2024-12-01 12:20:00	20	4.0
2	2024-12-02 20:10:00	40	4.8
3	2024-12-02 08:00:00	30	4.2
4	2024-12-03 13:15:00	15	4.7

	Order ID	User ID	Order Date	Meal Type	Dish Name	Order Status	
0	1001	U001	2024-12-01	Dinner	Spaghetti	Completed	
1	1002	U002	2024-12-01	Lunch	Caesar Salad	Completed	
2	1003	U003	2024-12-02	Dinner	Grilled Chicken	Canceled	
3	1004	U001	2024-12-02	Breakfast	Pancakes	Completed	
4	1005	U004	2024-12-03	Lunch	Caesar Salad	Completed	

	Amount (USD)	Time of Day	Rating	Session ID
0	15.0	Night	5.0	S001
1	10.0	Day	4.0	S002
2	12.5	Night	NaN	S003
3	8.0	Morning	4.0	S004
4	9.0	Day	4.0	S005

Step 2:

We used the `drop_duplicates()` function to eliminate any duplicate entries from the `user_details`, `cooking_sessions`, and `order_details` DataFrames in order to guarantee data quality and correctness. To increase the effectiveness of later analytic phases and to get rid of any biases or mistakes that can result from duplicate information, this step is essential.

Step 3:

We filled in missing values by using the median for numerical columns and 'Unknown' for categorical columns to assure data quality and get the data ready for analysis. Additionally, to provide precise time-based analysis, we transformed the 'Order Date' and 'Session Start' columns to the datetime format. Lastly, we examined each Data Frame's column names to confirm their structure and spot any possible irregularities.

```
➦ Index(['Order ID', 'User ID', 'Order Date', 'Meal Type', 'Dish Name',  
        'Order Status', 'Amount (USD)', 'Time of Day', 'Rating', 'Session ID'],  
        dtype='object')  
Index(['Session ID', 'User ID', 'Dish Name', 'Meal Type', 'Session Start',  
        'Session End', 'Duration (mins)', 'Session Rating'],  
        dtype='object')  
Index(['User ID', 'User Name', 'Age', 'Location', 'Registration Date', 'Phone',  
        'Email', 'Favorite Meal', 'Total Orders'],  
        dtype='object')
```

Step 4:

We standardised the column names in all Data Frames by eliminating any leading or following whitespace and changing them to lowercase in order to guarantee readability and uniformity. This stage increases the readability of the code and lowers the possibility of mistakes in further examination.

```
➦ Index(['order id', 'user id', 'order date', 'meal type', 'dish name',  
        'order status', 'amount (usd)', 'time of day', 'rating', 'session id'],  
        dtype='object')  
Index(['session id', 'user id', 'dish name', 'meal type', 'session start',  
        'session end', 'duration (mins)', 'session rating'],  
        dtype='object')  
Index(['user id', 'user name', 'age', 'location', 'registration date', 'phone',  
        'email', 'favorite meal', 'total orders'],  
        dtype='object')
```

2. Data Merging:

We combined the `order_details`, `cooking_sessions`, and `user_details` datasets in order to do the analysis. `cooking_sessions` contained information about cooking sessions, including `session_id`, `user_id`, `session_start`, `session_end`, and `session_duration`; `order_details` included order-related data, including `order_id`, `user_id`, `order_date`, and `order_amount`; and `user_details` included demographic data, including `user_id`, `age`, `location`, and `favorite_meal`.

Using `session_id` and `user_id` as common keys, we combined the datasets. To make sure that only rows with matching session identifiers were included, we first used the `session_id` to execute an inner join between `order_details` and `cooking_sessions`. To include demographic data, we next combined the resultant dataset with `user_details` on `user_id`. Only records with valid and matching identifiers were kept thanks to the use of inner joins.

We renamed columns with names that were similar across datasets to increase clarity. To prevent misunderstanding, for instance, we renamed columns such as `user_id_x` and `user_id_y` to `order_user_id` and `user_user_id`. By doing this step, the dataset was made easier to read and understand.

Following the merging, we checked the final dataset's structure to make sure all required columns were there and that data types were appropriately assigned, especially for numeric and date-related columns. This stage made sure that there were no crucial values missing and that the combined data was correct and clean.

	order_id	user_id	order_date	meal_type_x	dish_name_x	order_status	\
0	1001	U001	2024-12-01	Dinner	Spaghetti	Completed	
1	1002	U002	2024-12-01	Lunch	Caesar Salad	Completed	
2	1003	U003	2024-12-02	Dinner	Grilled Chicken	Canceled	
3	1004	U001	2024-12-02	Breakfast	Pancakes	Completed	
4	1005	U004	2024-12-03	Lunch	Caesar Salad	Completed	
	amount (usd)	time of day	rating	session_id	...	duration (mins)	\
0	15.0	Night	5.0	S001	...	30	
1	10.0	Day	4.0	S002	...	20	
2	12.5	Night	NaN	S003	...	40	
3	8.0	Morning	4.0	S004	...	30	
4	9.0	Day	4.0	S005	...	15	
	session_rating	user_name	age	location	registration_date	\	
0	4.5	Alice Johnson	28	New York	2023-01-15		
1	4.0	Bob Smith	35	Los Angeles	2023-02-20		
2	4.8	Charlie Lee	42	Chicago	2023-03-10		
3	4.2	Alice Johnson	28	New York	2023-01-15		
4	4.7	David Brown	27	San Francisco	2023-04-05		
	phone	email	favorite_meal	total_orders			
0	123-456-7890	alice@email.com	Dinner	12			
1	987-654-3210	bob@email.com	Lunch	8			
2	555-123-4567	charlie@email.com	Breakfast	15			
3	123-456-7890	alice@email.com	Dinner	12			
4	444-333-2222	david@email.com	Dinner	10			

[5 rows x 25 columns]

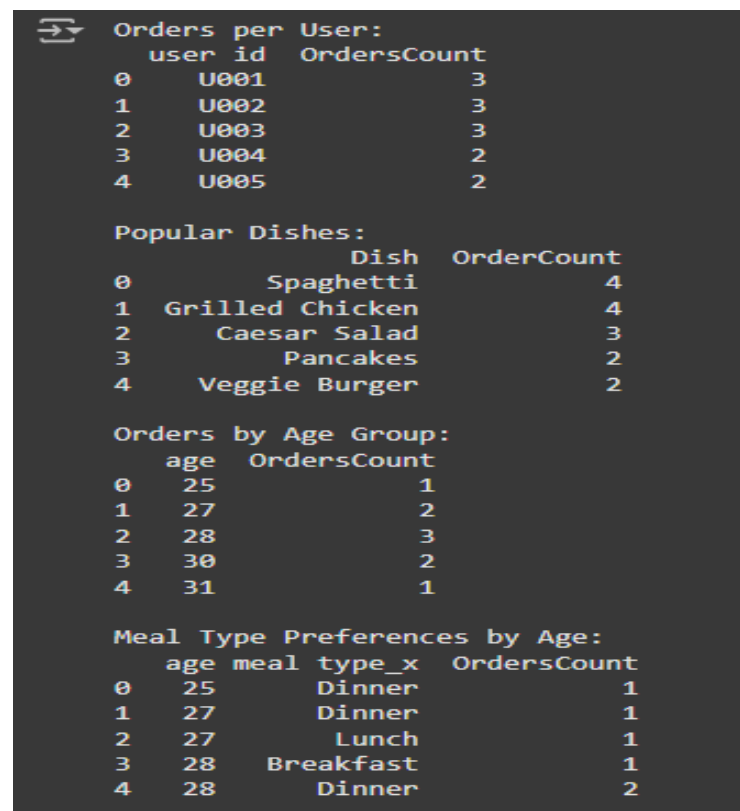
3. Data Analysis:

We concentrated on three main areas to better understand user behaviour: the quantity of orders placed by each user, the most popular dishes, and the demographic variables affecting order trends. Finding trends that could guide business tactics like menu optimisation and personalised marketing was made possible thanks in large part to these investigations.

We started by looking at the quantity of orders placed by each user. We determined how frequently each user places an order by classifying the combined dataset by `user_id` and counting the number of unique orders (`order_id`). High-order users who would be suitable candidates for targeted promotions or loyalty programs were identified by this analysis. On the other hand, re-engagement tactics may target consumers who place less orders.

We then concentrated on determining the most popular foods by looking at how often they were ordered. We determined the best-performing meals by calculating the frequency of orders for each dish throughout the dataset. These dishes might then be featured more prominently on the menu or in marketing efforts. Businesses can use their kitchen and inventory resources by knowing which meals are popular, guaranteeing that popular items are always available and ready to serve.

Finally, we investigated how demographic characteristics especially, age affect meal choices and order trends. We were able to learn more about how different age groups like meal kinds or dishes by classifying users according to their age and examining their ordering patterns. Businesses may improve customer happiness and perhaps increase revenue by using this data to help them customise their products to fit the demands of demographic groups. For instance, although older users might enjoy traditional meals, younger users could appreciate more creative or modern foods.



```
➡ Orders per User:
  user_id  OrdersCount
0      U001           3
1      U002           3
2      U003           3
3      U004           2
4      U005           2

Popular Dishes:
      Dish  OrderCount
0    Spaghetti           4
1  Grilled Chicken           4
2   Caesar Salad           3
3     Pancakes           2
4  Veggie Burger           2

Orders by Age Group:
   age  OrdersCount
0    25             1
1    27             2
2    28             3
3    30             2
4    31             1

Meal Type Preferences by Age:
   age meal_type_x  OrdersCount
0    25      Dinner           1
1    27      Dinner           1
2    27      Lunch            1
3    28  Breakfast            1
4    28      Dinner           2
```

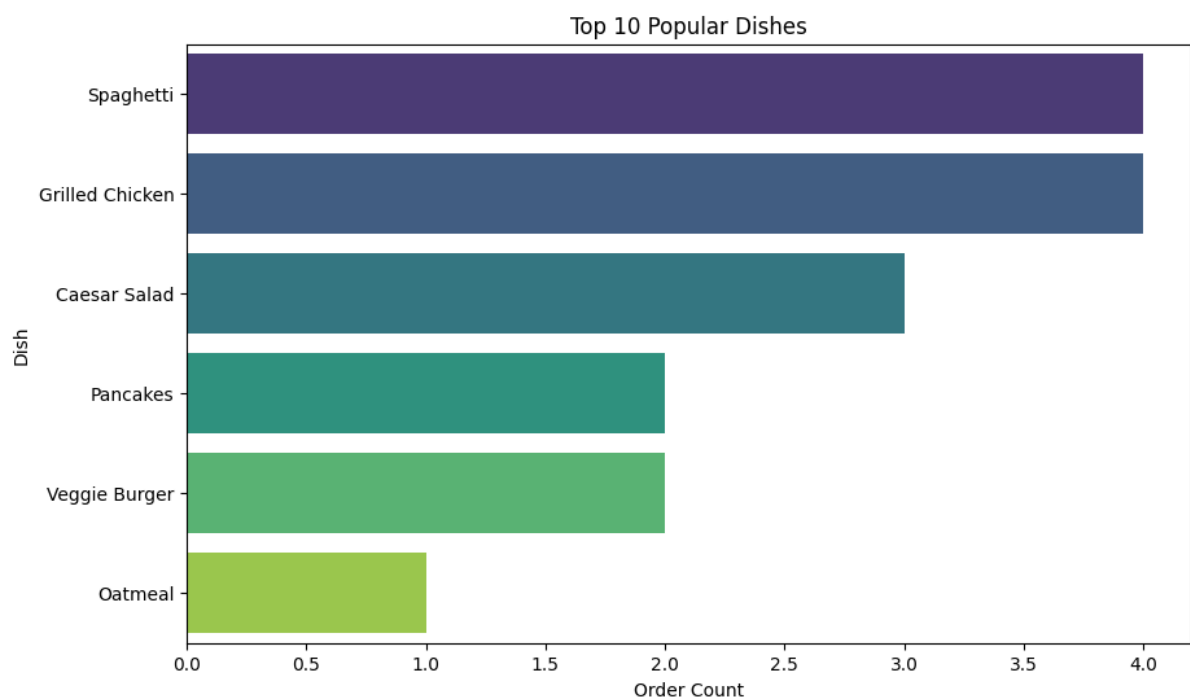
4. Visualizations:

a. "Top 10 Popular Dishes."

Key Observations:

- With about four orders, spaghetti is the most popular meal.
- Grilled Chicken comes in second place, with an order count slightly lower than Spaghetti.
- Caesar Salad ranks third, followed by Pancakes and Veggie Burger.
- Of the top 10, oatmeal is the least popular item with about one order, followed by pancakes and veggie burgers.

The top ten most ordered dishes are displayed in this bar chart. Of the top 10, **spaghetti and grilled chicken** stand out as the obvious choices, while **oatmeal** seems to be the least preferred. Marketing plans and menu planning might benefit from this study, which offers insightful information about consumer preferences.

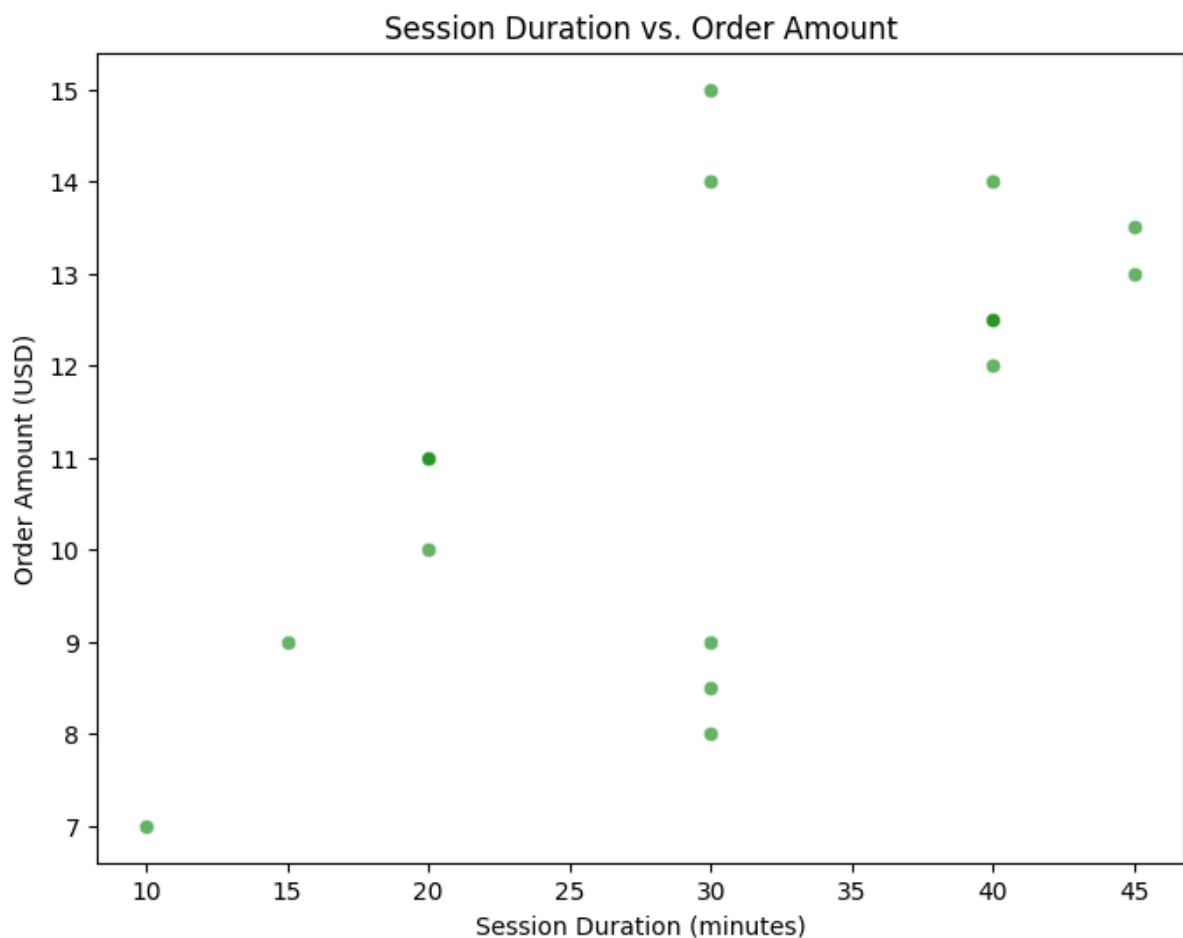


b. "Session Duration vs. Order Amount."

Key Observations:

- No discernible trend: The length of the session and the quantity of orders don't seem to be strongly or consistently correlated. The points lack a distinct upward or descending trend and are dispersed.
- Data points are clustered: A few data point clusters are seen. For instance, order quantities ranging from 8 to 14 USD are clustered around the 30-minute mark.
- Outliers: A small number of data points differ from the primary clusters. One data point, for example, has an order amount of about \$7 USD and a session time of about 10 minutes.

The link between order quantity and session time is shown in this scatter plot. There is no discernible relationship between these two variables, according to the graphic. The data points are dispersed and lack a clear pattern, indicating that the **length of a cooking session does not always affect the quantity ordered.**

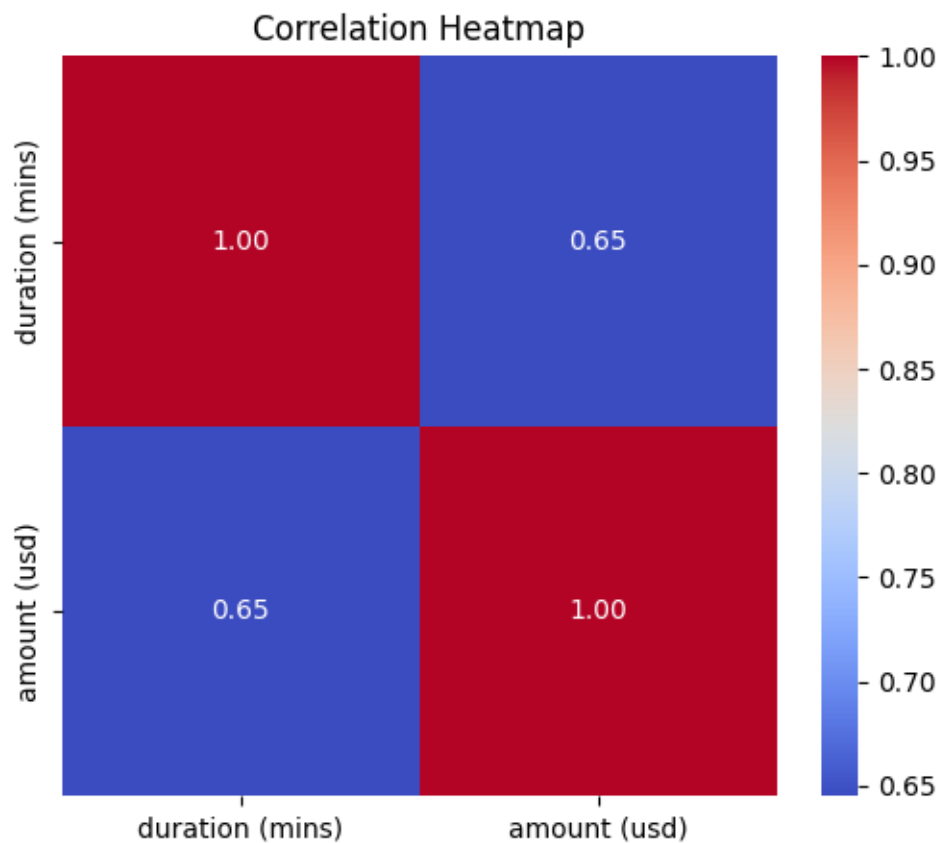


c. "Correlation Heatmap of duration (mins)" and "amount (usd)"

Key Observations:

- **Positive Correlation:** The heatmap demonstrates a robust positive relationship between order quantity and session time. A strong positive association is shown by the colour red.
- **Correlation Strength:** The correlation coefficient is represented by the value inside each cell. The correlation coefficient in this instance is 0.65, indicating a moderately to strongly favourable link.

There is a significant **positive link between session length and order quantity**, according to this correlation heatmap. The order quantity often rises in tandem with the length of the session. This implies that higher order values are linked to longer cooking periods.

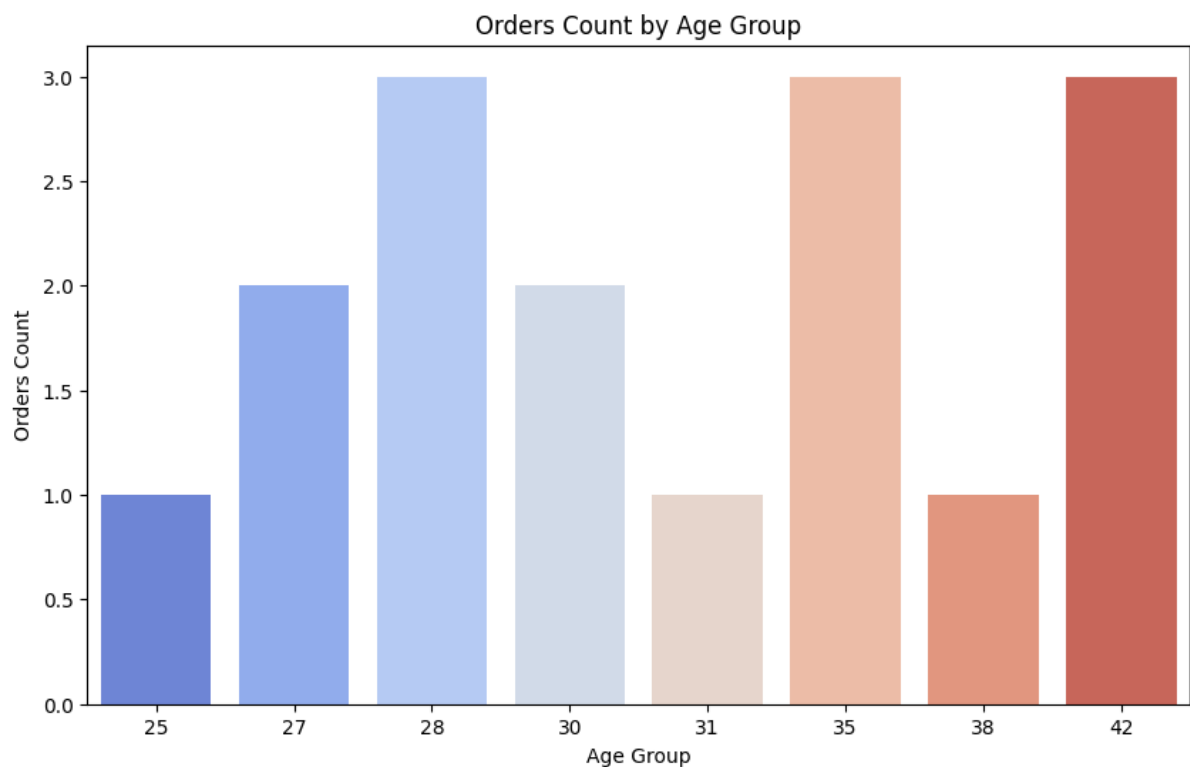


d. "Orders Count by Age Group."

Key Observations:

- Age 28 has the largest order count: Users in this age group placed the most orders, as indicated by the tallest bar for this age group.
- The lowest order counts are found in those aged 25 and 31: The fact that the age groups 25 and 31 have the smallest bars indicates that, in comparison to other age groups, consumers in these groups placed less orders.
- Overall trend: The order count for each age group shows no discernible rising or declining trend. It appears that the number of orders varies.

The distribution of orders among various age groups is seen in this bar chart. According to the findings, users in the age categories of **25 and 31 placed the fewest orders**, while those in the age group of **28 placed the most**. There are variations seen between age groups, indicating that the general trend in order count is not constant throughout age groups.



5. Conclusion

The purpose of this investigation was to look into user preferences and behaviour regarding cooking sessions and order patterns. Preprocessing, combining, and exploratory examination of the data revealed numerous important insights:

- **User Ordering Patterns:**

- o A moderately positive association between session length and order quantity was found through analysis, indicating that longer cooking sessions often provide greater order values.

- o The age group 28 users placed the most orders, whilst the age groups 25 and 31 users placed the fewest orders.

- **Popular Dishes:**

- o The most well-liked dishes among patrons were grilled chicken and spaghetti.

- o Pancakes, veggie burgers, and Caesar salad also shown a high level of interest. These findings provide valuable insights into user behaviour and preferences.

The following business strategies can be optimised by using this information:

- Marketing initiatives that are tailored to certain age groups and consumer interests are known as targeted marketing.

- Menu Planning: Improving menu selections according to customer preferences and well-liked foods.

- Session Optimisation: Determining the variables that affect session length and how they affect order value.

The association between user demographics, session length, and order quantity might be further examined. Additional information on user behaviour and preferences may be obtained by looking at variables like the time of day, day of the week, and seasonal tendencies.