About Dataset

This is dummy data, that I have generated by using the 'NumPy' Library of Python. This data shows how much a user spends time on their devices using Social Media.I generated this data to train an AI model for myself for practice purposes only.The description for each column is as follows:

- age: The age of the user.
- gender: The gender identity of the user (Male, Female, Non-binary).
- demographics: The type of area the user resides in (Urban, Suburban, Rural).
- interests: The user's primary area of interest or hobby.
- device_type: The type of device used by the user (Mobile).
- location:The country of residence for the user.
- platform: The social media platform where the user spends time.
- profession: The user's occupation or professional status.
- income: The yearly income of the user.
- indebt: Indicates whether the user is in debt (True or False).
- homeowner: Indicates whether the user owns a home (True or False).
- owns_cars: Indicates whether the user owns cars (True or False).

In [1]:
```
# Import the required libraries for data analysis process
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
import warnings
import plotly.offline as pyo
warnings.filterwarnings('ignore')
```

In [2]:
```
# Load the data using the pandas read function
data=pd.read_csv('C://Users/vinod//Downloads//archive//dummy_data.csv')
data.head()
```

Out[2]:

	age	gender	time_spent	platform	interests	location	demographics	profession	income	in
0	56	male	3	Instagram	Sports	United Kingdom	Urban	Software Engineer	19774	
1	46	female	2	Facebook	Travel	United Kingdom	Urban	Student	10564	
2	32	male	8	Instagram	Sports	Australia	Sub_Urban	Marketer Manager	13258	F
3	60	non-binary	5	Instagram	Travel	United Kingdom	Urban	Student	12500	F
4	25	male	1	Instagram	Lifestlye	Australia	Urban	Software Engineer	14566	F

Data Preprocessing steps

- Checking the data shape
- Checking the null values
- Some Stastical information
- information about data

In [3]:
```
def checking_data_information(data_frame):
    shape=data_frame.shape
    print('*'*80)
    null_values=data_frame.isna().sum()/len(data_frame)
    print('$'*80)
    data_frame_info=data_frame.info()
    print('&'*80)
    duplicated_value=data.duplicated().sum()
    print(f'The data set contains {duplicated_value} values')
    return shape,null_values,data_frame_info
print(checking_data_information(data))
```

```
********************************************************************************
***
$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$
$$$
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   age           1000 non-null   int64
 1   gender        1000 non-null   object
 2   time_spent    1000 non-null   int64
 3   platform      1000 non-null   object
 4   interests     1000 non-null   object
 5   location      1000 non-null   object
 6   demographics  1000 non-null   object
 7   profession    1000 non-null   object
 8   income        1000 non-null   int64
 9   indebt        1000 non-null   bool
 10  isHomeOwner   1000 non-null   bool
 11  Owns_Car      1000 non-null   bool
dtypes: bool(3), int64(3), object(6)
memory usage: 73.4+ KB
&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&
&&&
The data set contains 0 values
((1000, 12), age             0.0
gender          0.0
time_spent      0.0
platform        0.0
interests       0.0
location        0.0
demographics    0.0
profession      0.0
income          0.0
indebt          0.0
isHomeOwner     0.0
Owns_Car        0.0
dtype: float64, None)
```

In [4]:
```
# Some statstical information about the data
data.describe().style.background_gradient(cmap='RdGy_r')
```

Out[4]:

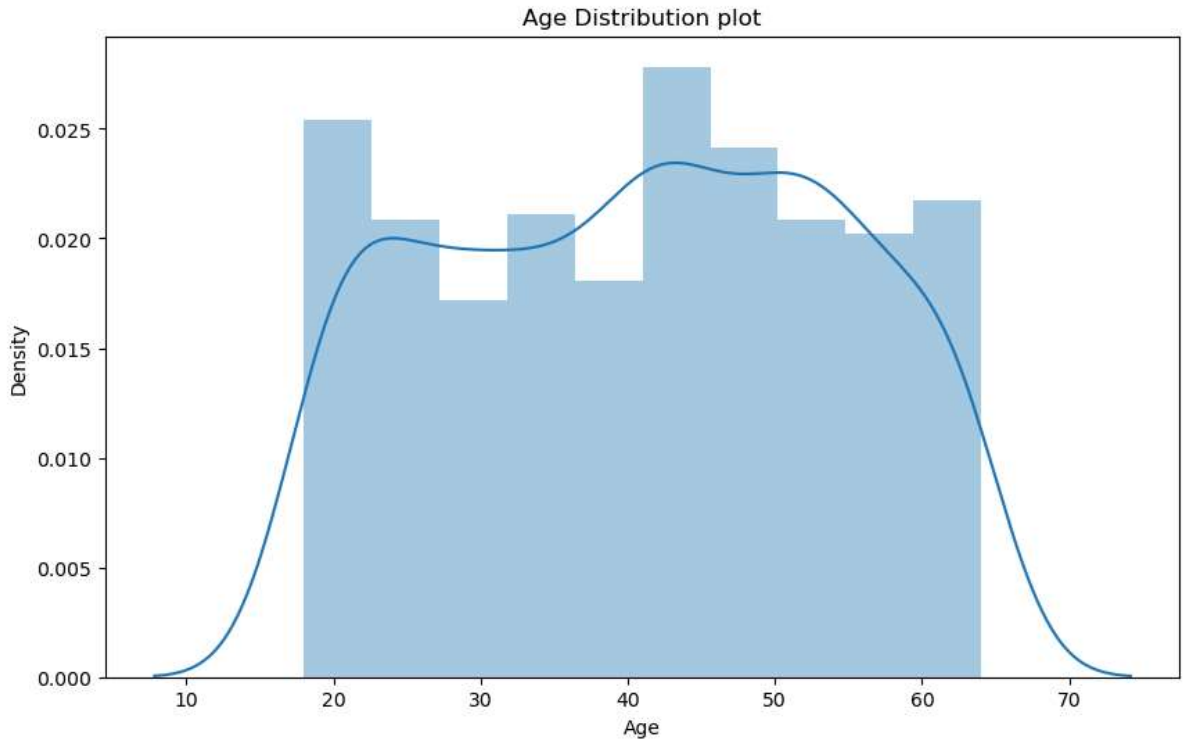
	age	time_spent	income
count	1000.000000	1000.000000	1000.000000
mean	40.986000	5.029000	15014.823000
std	13.497852	2.537834	2958.628221
min	18.000000	1.000000	10012.000000
25%	29.000000	3.000000	12402.250000
50%	42.000000	5.000000	14904.500000
75%	52.000000	7.000000	17674.250000
max	64.000000	9.000000	19980.000000

Explore Data Analysis Process

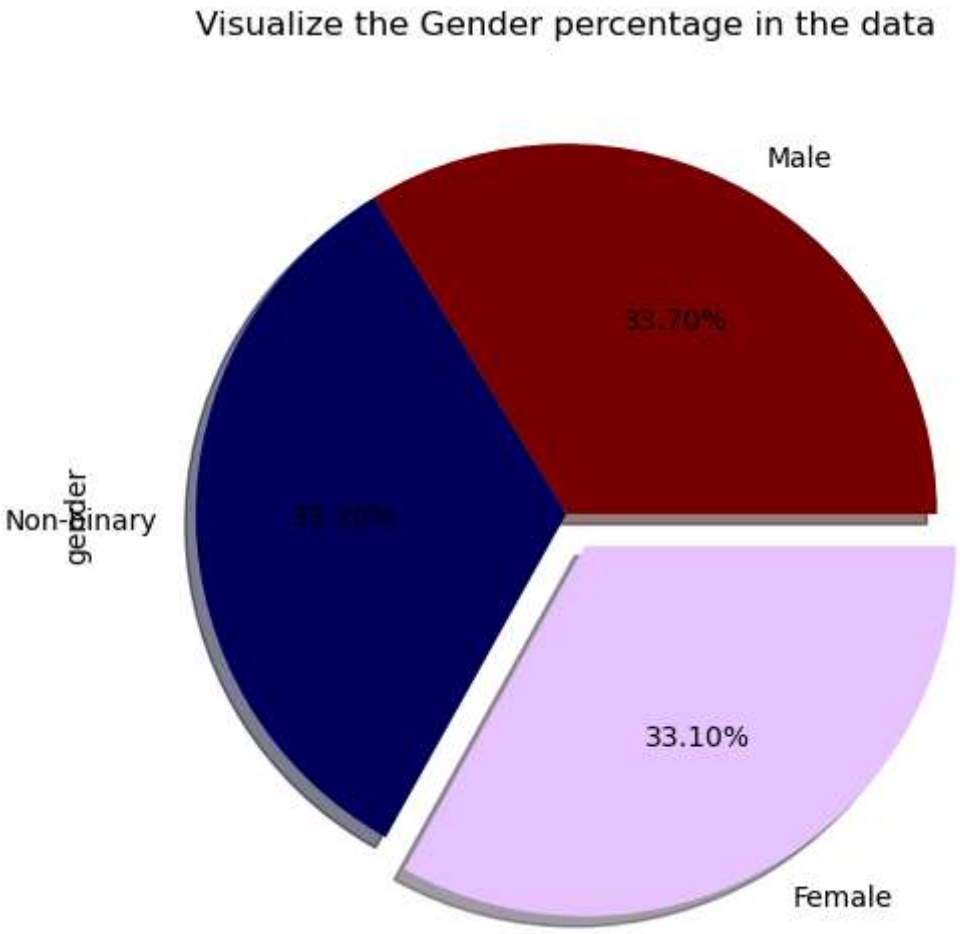
Questions Asked from the Data. We ask some find the some intresting information about the data

- Visualize the age distribution using a distplot.
- Create a pie chart to understand the gender percentage in the data.
- Visualize the unique plots in different locations.
- Visualize how many times students use different social media platforms in different locations.
- Create a data frame to show how much time each profession spends on different social media platforms on average.
- Create a data frame for demographics, platforms, and professions.

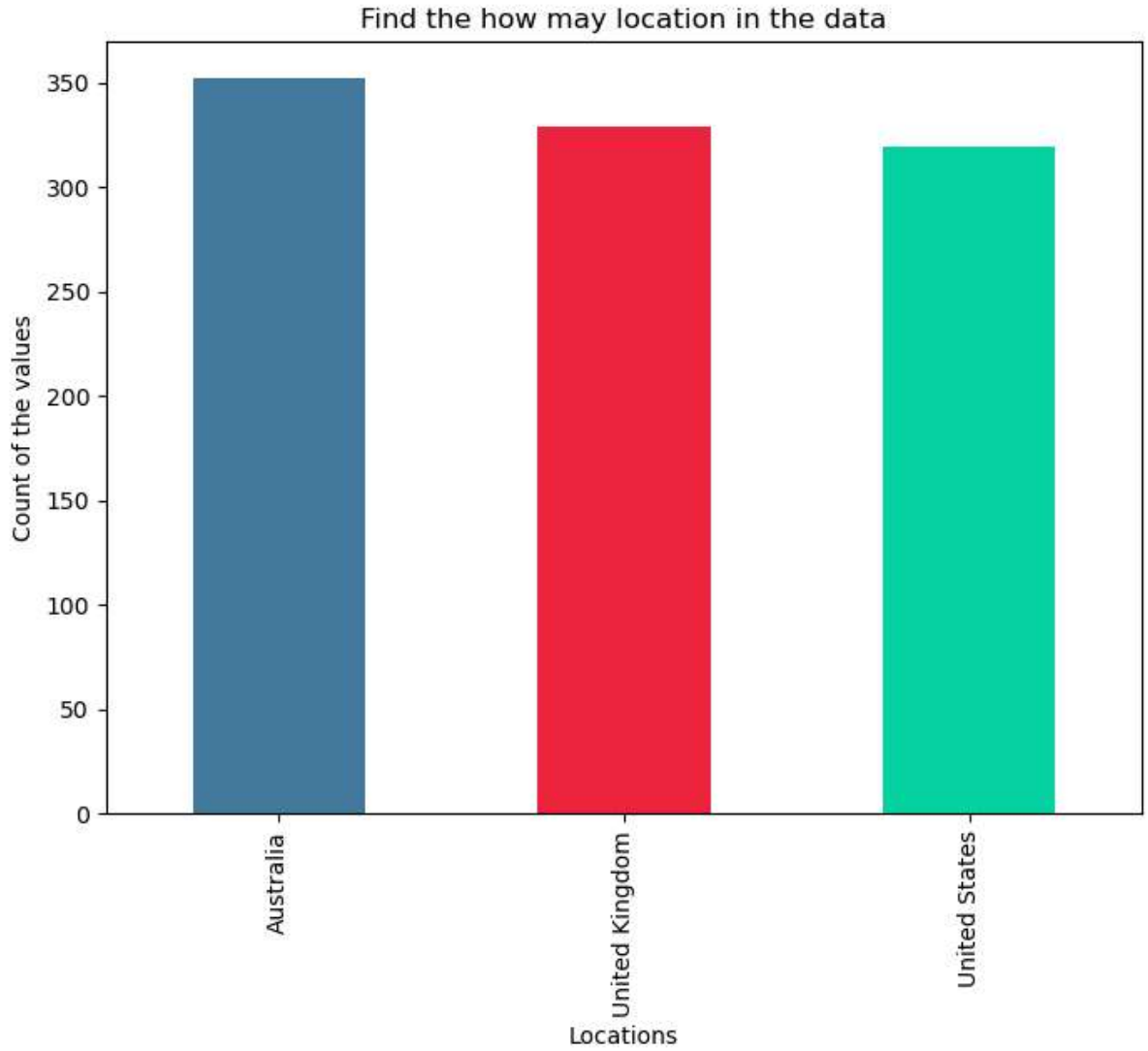
In [5]:
```
# Visualize the age distribution
plt.figure(figsize=(10,6))
sns.distplot(data['age'],hist=True,kde=True)
plt.title("Age Distribution plot")
plt.xlabel('Age')
plt.ylabel("Density")
plt.show()
```



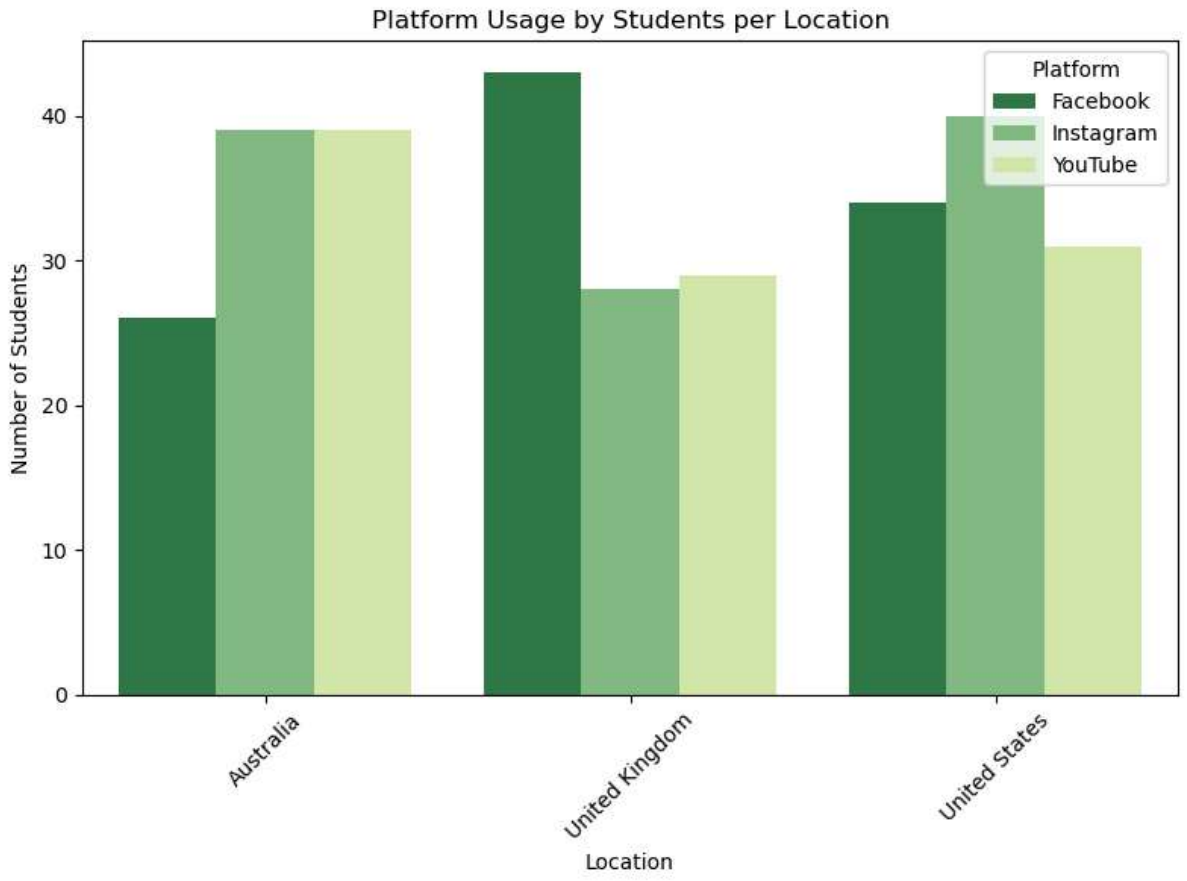
In [6]:
```
# Let's Visualize the Gender Percentage in the data
data['gender'].value_counts().sort_values(ascending=False)\
.plot(kind='pie',figsize=(8,6),
     explode=[0,0,0.1],
    labels=['Male','Non-binary','Female'],
    colors=['#780000','#03045e','#e7c6ff'],
    autopct='%1.2f%%',
    shadow=True)
plt.title("Visualize the Gender percentage in the data")
plt.show()
```



In [7]:
```
# Let;s find the How many location in the data
data['location'].value_counts().sort_values(ascending=False)\
.plot(kind='bar',figsize=(8,6),color=['#457b9d','#ef233c','#06d6a0'])
plt.title("Find the how may location in the data")
plt.xlabel('Locations')
plt.ylabel("Count of the values")
plt.show()
```



In [8]:
```
# Let's find some intresting question
'''Let's find the how many total student visit the social media plotform
in differnt locations
'''
student_data=data[data['profession']=='Student']
location_platform_data_frame=pd.DataFrame(student_data.groupby('location',)['pl
# Reshaping the DataFrame for seaborn barplot
platform_counts = location_platform_data_frame.reset_index().melt(id_vars='loca
# Plotting with seaborn
plt.figure(figsize=(8, 6))
sns.barplot(data=platform_counts, x='location', y='count', hue='platform', pale
# Adding labels and title
plt.xlabel('Location')
plt.ylabel('Number of Students')
plt.title('Platform Usage by Students per Location')

# Display the plot
plt.legend(title='Platform')
plt.xticks(rotation=45)  # Rotate x-axis labels for better readability
plt.tight_layout()  # Adjust layout to prevent clipping of labels
plt.show()
```



In [9]:
```
'''
Let's find the average time for the all profession employess
spends on the each social media platfom and
'''
profession_platform=pd.DataFrame(data.groupby('profession')['platform'].value_
profession_time=pd.DataFrame(data.groupby('profession')['time_spent'].mean())
all_information=pd.merge(profession_platform,profession_time,on='profession')
all_information.style.background_gradient(cmap='ocean')
```

Out[9]:

profession	Facebook	Instagram	YouTube	time_spent
Marketer Manager	110	128	117	5.095775
Software Engineer	94	128	114	4.949405
Student	103	107	99	5.038835

In [10]:
```
'''Create a chart to understanding the how many people mostly
using the social media with different demographics we do groupby function and t
we do sort the values
'''
grouping_data=pd.DataFrame(data.groupby(['platform','demographics'])['professi
grouping_data.style.background_gradient(cmap='Reds')
```

Out[10]:

	profession	Marketer Manager	Software Engineer	Student
platform	demographics			
Facebook	Rural	31	33	32
	Sub_Urban	40	34	32
	Urban	39	27	39
Instagram	Rural	51	44	41
	Sub_Urban	47	40	31
	Urban	30	44	35
YouTube	Rural	30	42	36
	Sub_Urban	45	37	29
	Urban	42	35	34

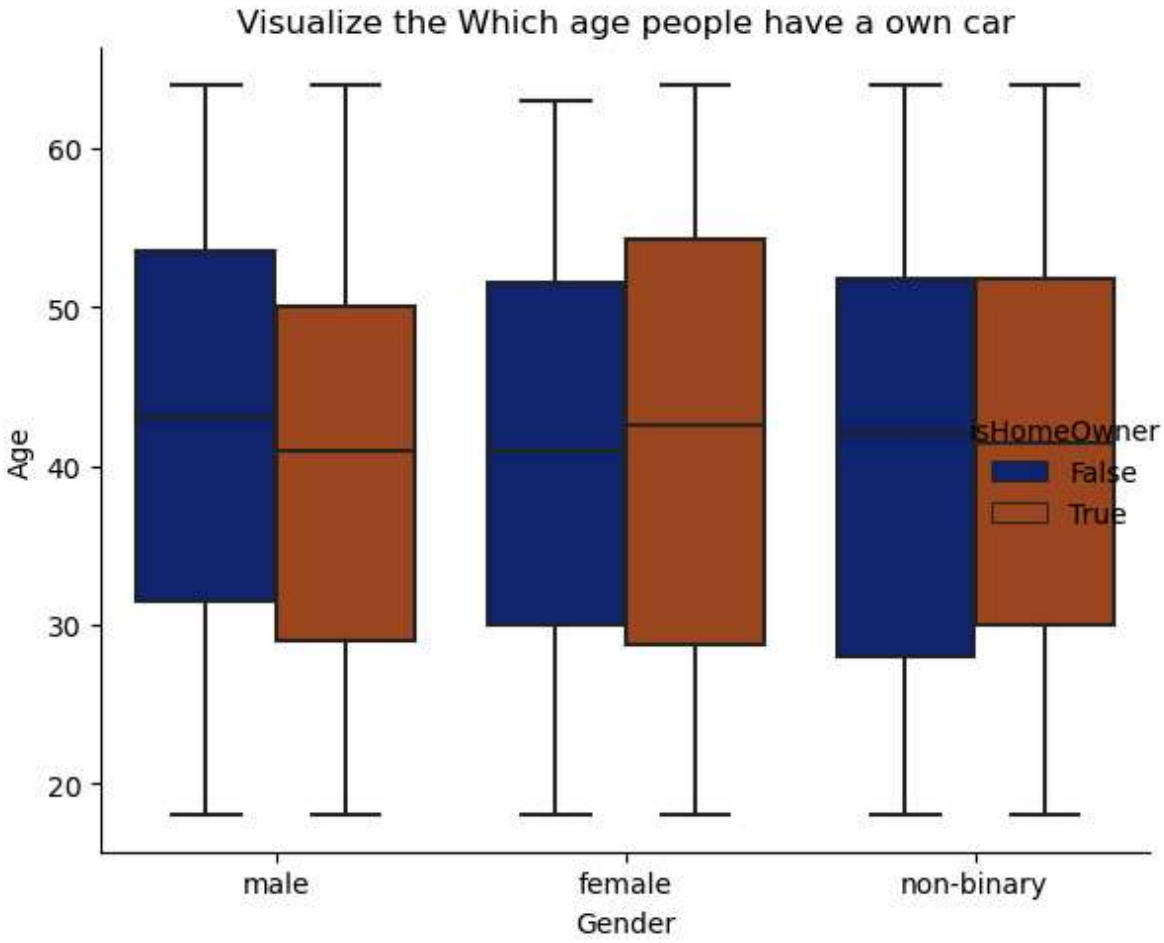
Based on the data, we can make the following observations:

- The age distribution shows that most of the people fall in the age range of 40 to 55.
- Looking at the gender category, we can see that males dominate the data, followed by non-binary genders.
- In terms of unique values, Australia has the highest values, followed by the UK and the USA.

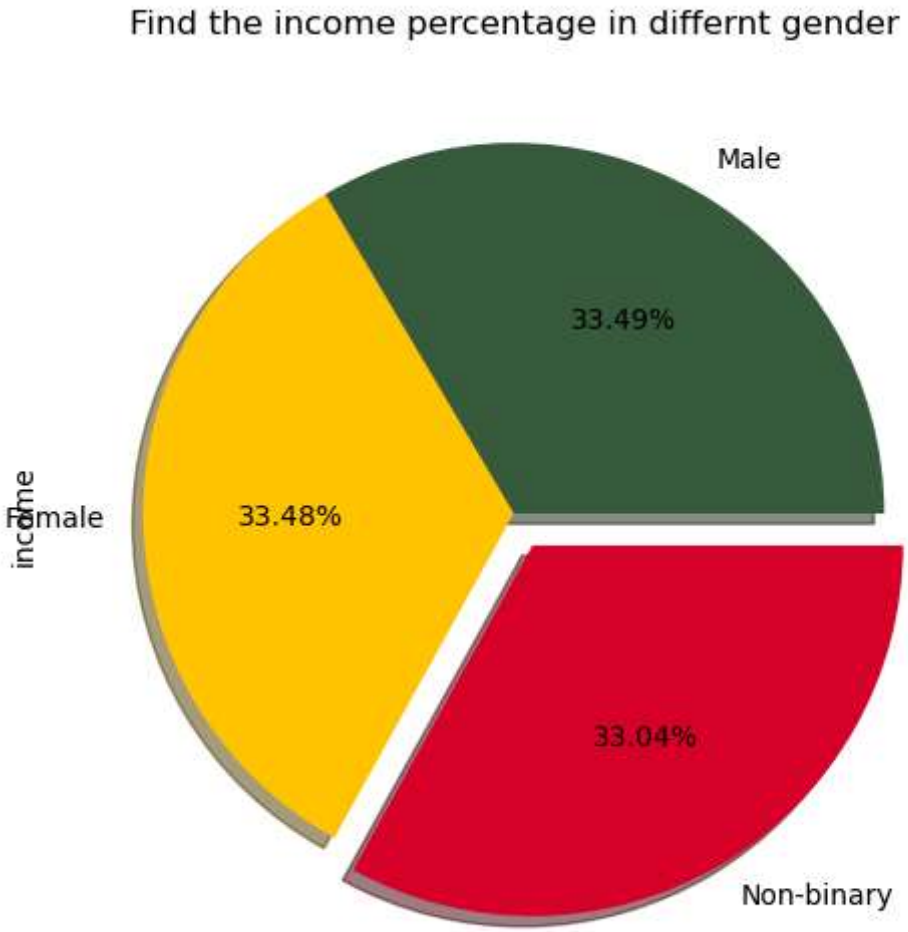
- When analyzing the social media behavior of students in different locations, we observed that in Australia, the most used platforms are YouTube and Instagram, whereas in the UK, it's Facebook and YouTube, and in the USA, it's Instagram and Facebook.
- We also observed that software engineers spend less time on social media compared to

In [11]:
```
'''
Create a boxplot to understand the how many people have
own car in differnt age with differnt gender using the boxplots

'''
plt.figure(figsize=(10,6))
sns.catplot(data=data,x='gender',y='age',hue='isHomeOwner',kind='box',palette=
plt.title("Visualize the Which age people have a own car")
plt.xlabel('Gender')
plt.ylabel('Age')
plt.tight_layout()
plt.show()
```

```
<Figure size 1000x600 with 0 Axes>
```



In [12]:
```
'''
Let's find the source of income of the different gender
using the pie chart
'''
data.groupby('gender')['income'].sum().sort_values(ascending=False)\
.plot(kind='pie',figsize=(8,6),
     explode=[0,0,0.1],
    labels=['Male','Female','Non-binary'],
    colors=['#3a5a40','#ffc300','#d90429'],
    autopct='%1.2f%%',
    shadow=True)
plt.title("Find the income percentage in differnt gender")
plt.show()
```

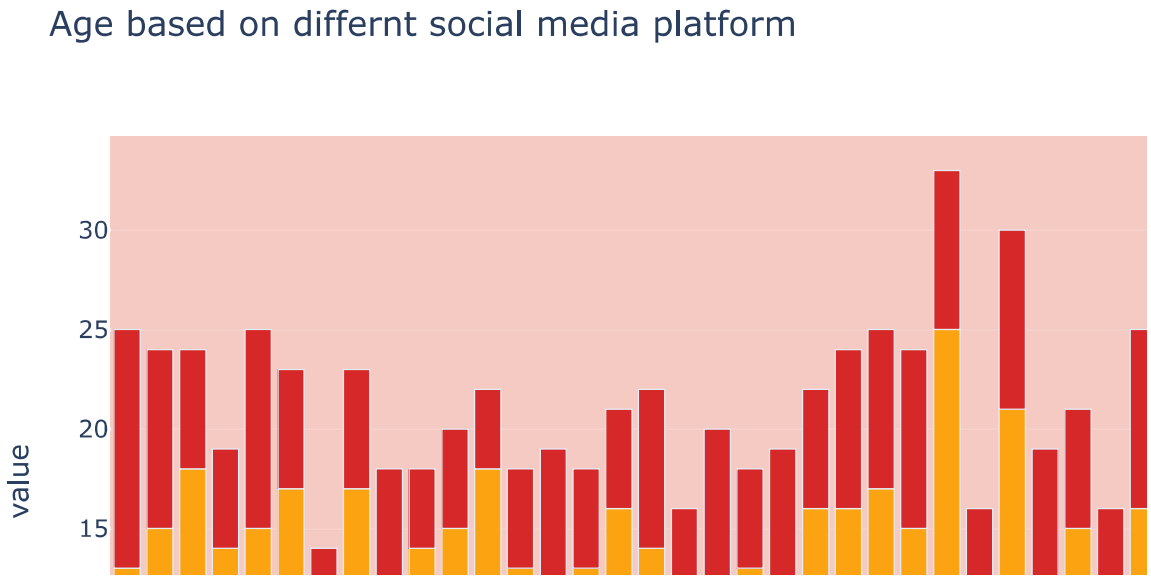


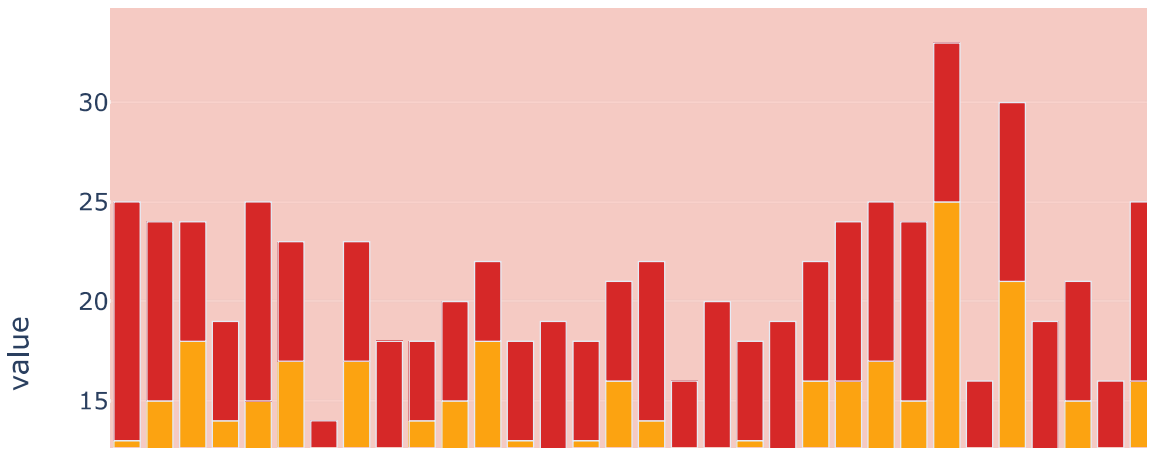
In [13]:
```
'''
find the average time the diffent profession spent on the instagram
and Facebook
'''
instagram=data[data['platform']=='Instagram']
instagram=pd.DataFrame(instagram.groupby('profession')['time_spent'].mean().so
face_book=data[data['platform']=='Facebook']
face_book_data=pd.DataFrame(face_book.groupby('profession')['time_spent'].mean
youtube=data[data['platform']=='YouTube']
youtube_info=pd.DataFrame(youtube.groupby('profession')['time_spent'].mean().s
final_data=pd.merge(instagram,face_book_data,on='profession',suffixes=('_insta
social_info=pd.merge(final_data,youtube_info,on='profession')
social_info.style.background_gradient(cmap='Wistia')
```

Out[13]:

profession	time_spent_instagram	time_spent_facebook	time_spent
Software Engineer	5.265625	4.797872	4.719298
Student	5.158879	5.048544	4.898990
Marketer Manager	5.031250	5.281818	4.991453

In [14]:
```
'''
Find the Which age pepole most used  social media platforms
in the data, we create a bar chart to understand the information clearly
'''
age_platform=data.groupby('age')['platform'].value_counts().sort_values(ascend
fig=px.bar(age_platform,x=age_platform.index,y=age_platform.columns,title='Age
fig.update_layout(legend_bgcolor='#e63946',plot_bgcolor='#f5cac3')
fig.show()
pyo.iplot(fig)
```

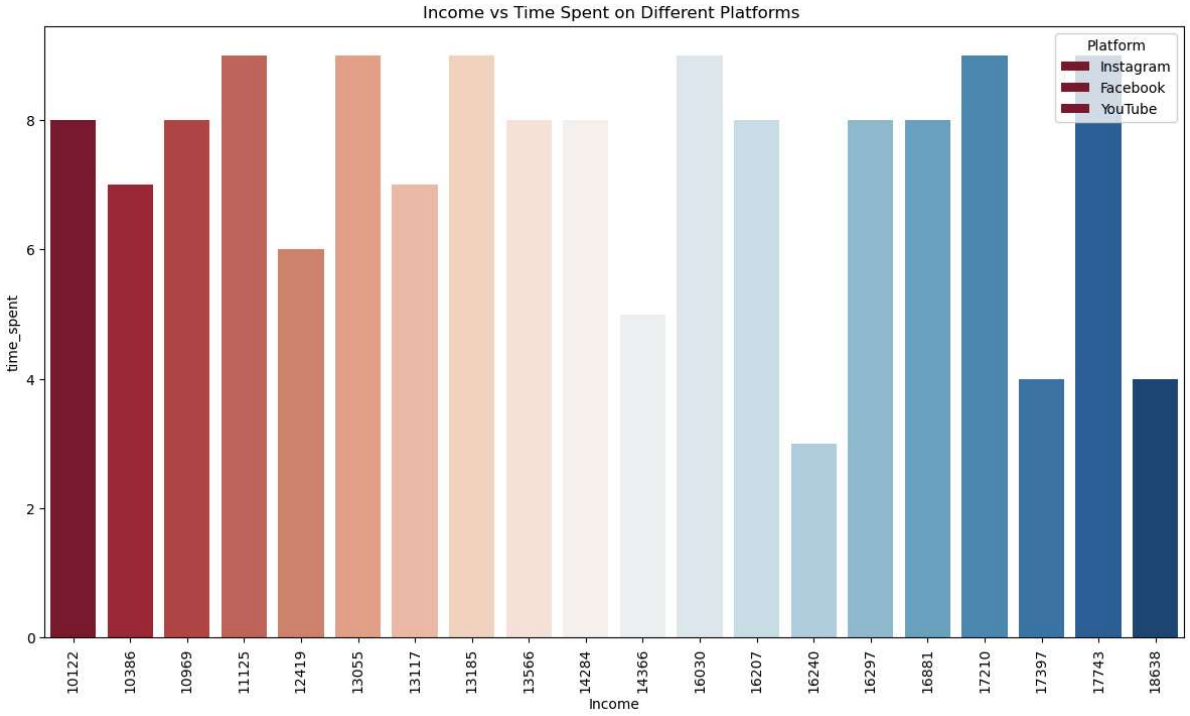



Age based on differnt social media platform
value
30
25
20
15

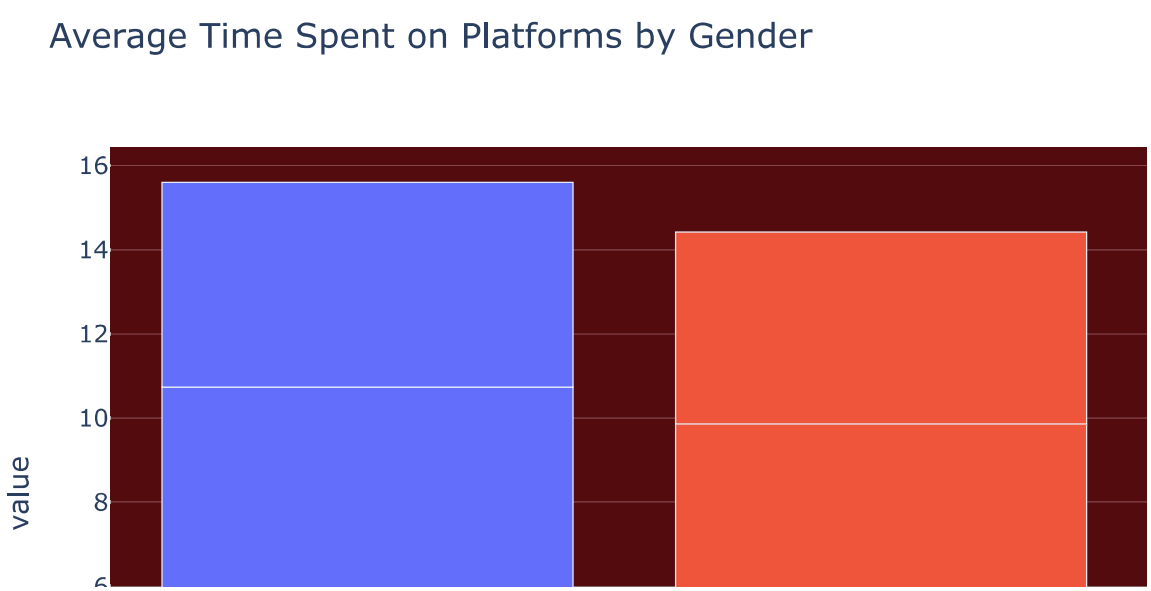
In [15]:
```
'''
We creat the barplot to understand the different intrests on the
differnet social media platfoms and we visualize with plotly libaries

'''
interest_platform=data.groupby('interests')['platform'].value_counts().sort_va
fig=px.histogram(interest_platform,x=interest_platform.index,y=interest_platfo
fig.update_layout(paper_bgcolor="#124559")
fig.show()
pyo.iplot(fig)
```

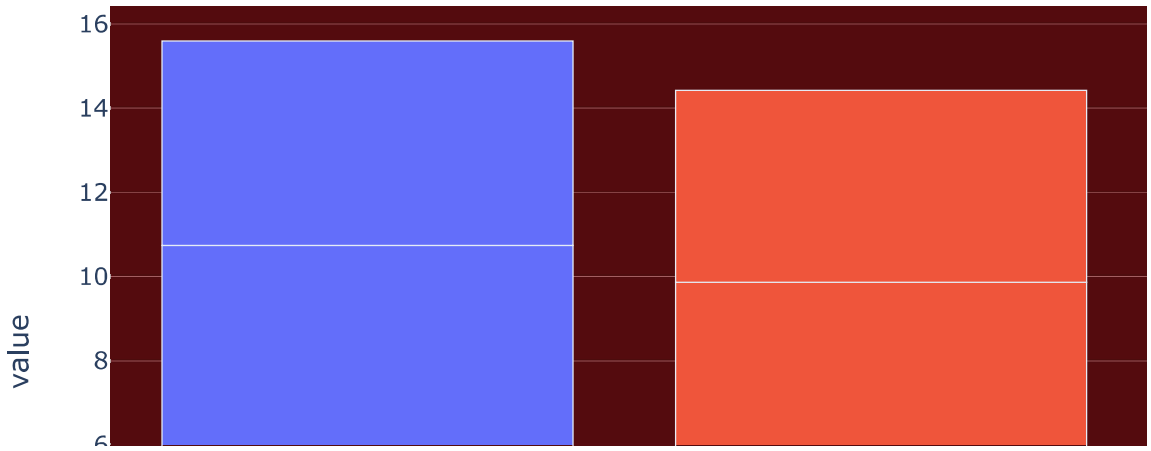

Platform preference in intrest wise
sum of value
120
100
80
60

In [16]:
```
'''
Find the top 20 income with time spent in differnt social media platforms
using the scatter plots
'''
plt.figure(figsize=(12,7))
for platform in data['platform'].unique():
    sns.barplot(x='income',y='time_spent',data=data[data['platform']==platform]
plt.title("Income vs Time Spent on Different Platforms ")
plt.xlabel('Income')
plt.ylabel('time_spent')
plt.tight_layout()
plt.legend(title='Platform')
plt.xticks(rotation=90)
plt.show()
```

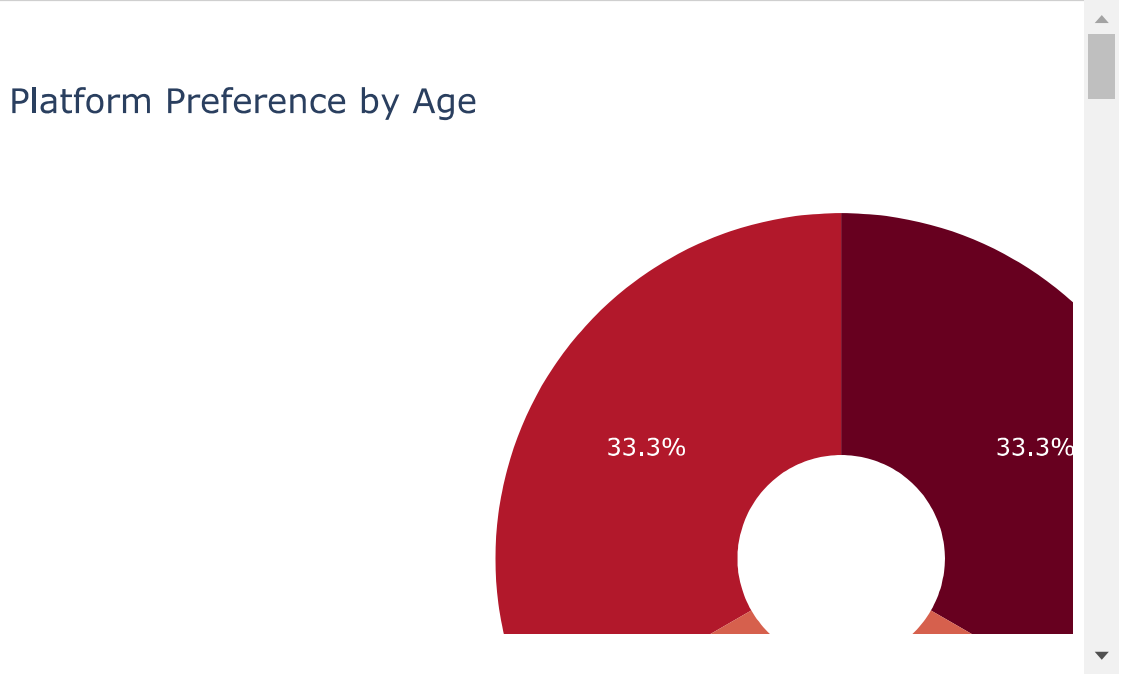


In [17]:
```
'''
# Let's find the average time spent on the each gender in differnt platforms
Fristly we groupby the columns and we combine them and we visualize it using th
bar chart
'''
age_time=pd.DataFrame(data.groupby(['gender','platform'])['time_spent'].mean()
fig=px.bar(age_time,x=age_time.index,y=age_time.columns,color=age_time.columns,
             labels={'x':'Gender', 'y':'Average Time Spent', 'color':'Platform
fig.update_layout(legend_bgcolor='#e63946',plot_bgcolor='#540b0e')
fig.show()
pyo.iplot(fig)
```

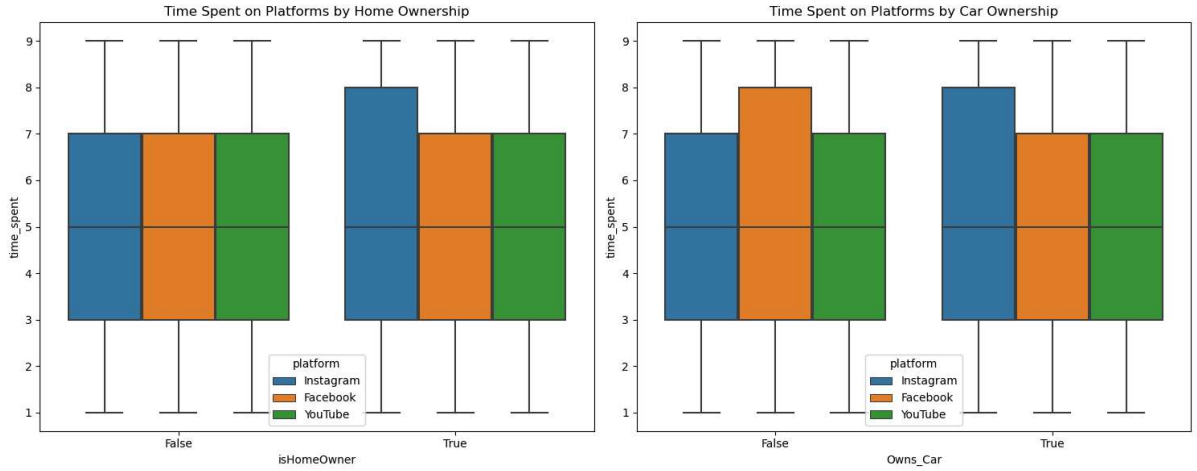



Average Time Spent on Platforms by Gender
value
16
14
12
10
8
6

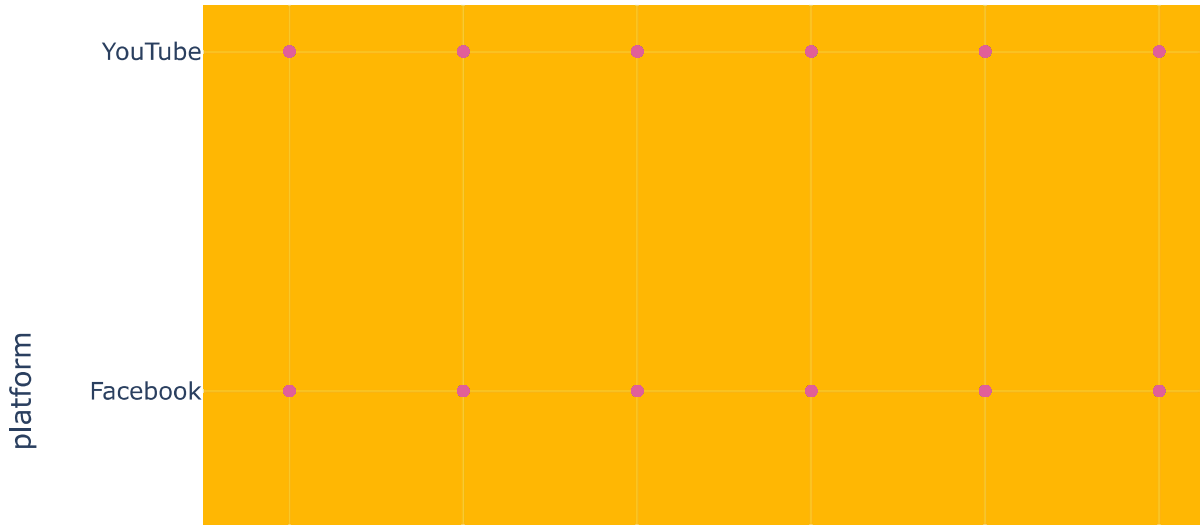
In [18]:
```
'''
Find the each platform in differnt category such as age and gender profession u
pie chart
'''
demographics = ['age', 'gender', 'profession']  # You can add more demographics

for demographic in demographics:
    # Calculate platform usage distribution for the demographic
    platform_info = data.groupby(demographic)['platform'].value_counts().unsta

    # Plotting the pie chart
    fig = px.pie(platform_info,
                 names=platform_info.columns,
                 title=f'Platform Preference by {demographic.capitalize()}',
                 labels={'platform':'Platform'},color_discrete_sequence=px.col
    fig.show()
    pyo.iplot(fig)
```

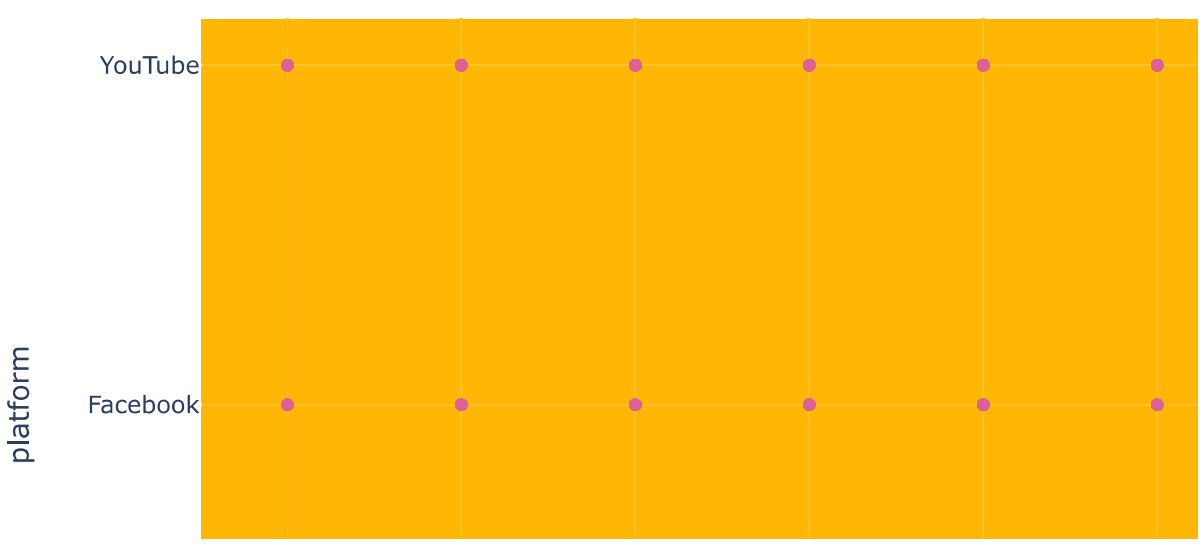


In [19]:
```
plt.figure(figsize=(15, 6))

# Boxplot for isHomeOwner
plt.subplot(1, 2, 1)
sns.boxplot(data=data, x='isHomeOwner', y='time_spent', hue='platform')
plt.title('Time Spent on Platforms by Home Ownership')

# Boxplot for Owns_Car
plt.subplot(1, 2, 2)
sns.boxplot(data=data, x='Owns_Car', y='time_spent', hue='platform')
plt.title('Time Spent on Platforms by Car Ownership')

plt.tight_layout()
plt.show()
```



In [20]:
```
# Let's find the how many people spend time on different platform in indebt
fig=px.scatter(data,x='time_spent',y='platform',color='indebt',color_discrete_
fig.update_layout(plot_bgcolor='#ffb703')
fig.show()
pyo.iplot(fig)
```




YouTube
Facebook
platform

In [21]:
```
# Create a bar chart to understand the time spent on differnt location in diffe
time_spent_platform=['platform','location']
for time_platform_location in time_spent_platform:
    time_spent_vs_platform_location=pd.DataFrame(data.groupby(time_platform_loc
    fig=px.bar(time_spent_vs_platform_location,color=time_spent_vs_platform_loc
               title='Time_spent Vs location in differnt platforms')
    fig.update_layout(plot_bgcolor='#e5e5e5')
    fig.show()
    pyo.iplot(fig)
```

In [ ]: