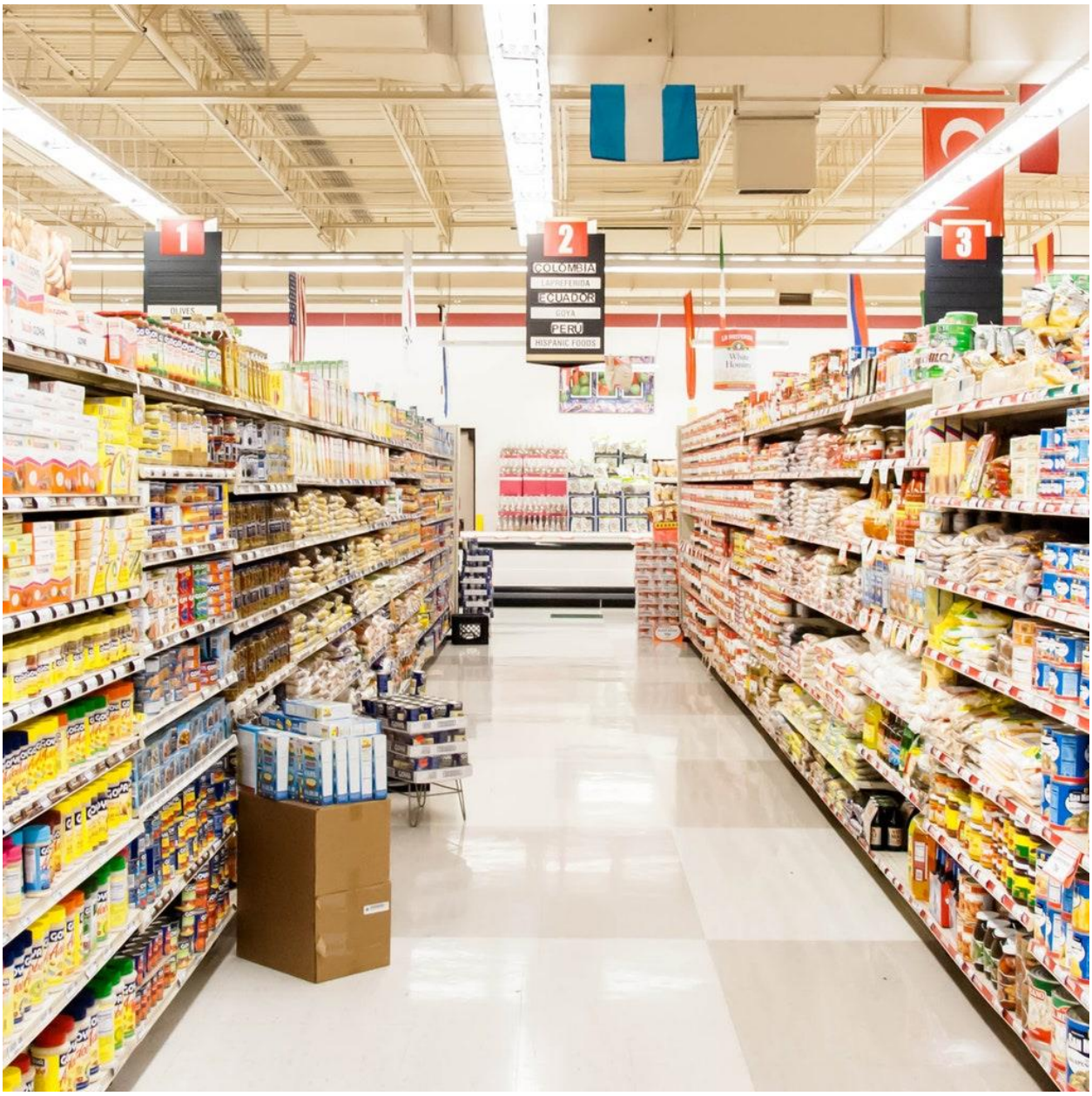
SUPERMARKET SALES DATA: A VISUALIZATION AND ANALYSIS

Supermarket adalah sebuah tempat dimana kita bisa membeli barang – barang kebutuhan kita yang dilakukan dengan sistem self-service.



Di artikel ini, saya ingin menganalisis data penjualan dari supermarket. Dataset yang digunakan diperoleh dari Kaggle dengan link berikut Supermarket sales | Kaggle.

Modul – modul yang digunakan dalam melakukan analisis antara lain Matplotlib, Pandas, Datetime, dan Seaborn. Pertama, dilakukan ***importing dataset*** seperti sebagai berikut:

```
In [2]: #Data importing
        dataset = pd.read_csv('supermarket_sales.csv')
```

Dataset supermarket sales:

```
In [4]: dataset
```

Out[4]:

	Invoice ID	Branch	City	Customer type	Gender	Product line	Unit price	Quantity	Tax 5%	Total	Date	Time	Payment	cogs	gross margin percentage	gross income
0	750-67-8428	A	Yangon	Member	Female	Health and beauty	74.69	7	26.1415	548.9715	1/5/2019	13:08	Ewallet	522.83	4.761905	26.1415
1	226-31-3081	C	Naypyitaw	Normal	Female	Electronic accessories	15.28	5	3.8200	80.2200	3/8/2019	10:29	Cash	76.40	4.761905	3.8200
2	631-41-3108	A	Yangon	Normal	Male	Home and lifestyle	46.33	7	16.2155	340.5255	3/3/2019	13:23	Credit card	324.31	4.761905	16.2155
3	123-19-1176	A	Yangon	Member	Male	Health and beauty	58.22	8	23.2880	489.0480	1/27/2019	20:33	Ewallet	465.76	4.761905	23.2880
4	373-73-7910	A	Yangon	Normal	Male	Sports and travel	86.31	7	30.2085	634.3785	2/8/2019	10:37	Ewallet	604.17	4.761905	30.2085
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
995	233-67-5758	C	Naypyitaw	Normal	Male	Health and beauty	40.35	1	2.0175	42.3675	1/29/2019	13:46	Ewallet	40.35	4.761905	2.0175
996	303-96-2227	B	Mandalay	Normal	Female	Home and lifestyle	97.38	10	48.6900	1022.4900	3/2/2019	17:16	Ewallet	973.80	4.761905	48.6900
997	727-02-1313	A	Yangon	Member	Male	Food and beverages	31.84	1	1.5920	33.4320	2/9/2019	13:22	Cash	31.84	4.761905	1.5920
998	347-56-2442	A	Yangon	Normal	Male	Home and lifestyle	65.82	1	3.2910	69.1110	2/22/2019	15:33	Cash	65.82	4.761905	3.2910
999	849-09-3807	A	Yangon	Member	Female	Fashion accessories	88.34	7	30.9190	649.2990	2/18/2019	13:28	Cash	618.38	4.761905	30.9190

1000 rows × 17 columns

(data ditampilkan tidak semuanya)

Nilai statistik dasar dari dataset supermarket sales tersebut adalah:

```
In [4]: dataset.describe()
```

Out[4]:

	Unit price	Quantity	Tax 5%	Total	cogs	gross margin percentage	gross income	Rating
count	1000.000000	1000.000000	1000.000000	1000.000000	1000.00000	1.000000e+03	1000.000000	1000.00000
mean	55.672130	5.510000	15.379369	322.966749	307.58738	4.761905e+00	15.379369	6.97270
std	26.494628	2.923431	11.708825	245.885335	234.17651	6.220360e-14	11.708825	1.71858
min	10.080000	1.000000	0.508500	10.678500	10.17000	4.761905e+00	0.508500	4.00000
25%	32.875000	3.000000	5.924875	124.422375	118.49750	4.761905e+00	5.924875	5.50000
50%	55.230000	5.000000	12.088000	253.848000	241.76000	4.761905e+00	12.088000	7.00000
75%	77.935000	8.000000	22.445250	471.350250	448.90500	4.761905e+00	22.445250	8.50000
max	99.960000	10.000000	49.650000	1042.650000	993.00000	4.761905e+00	49.650000	10.00000

Setelah diimport dan diketahui datasetnya seperti apa, selanjutnya dilakukan ***pengecekan adanya missing values***

```
In [3]: #Data Cleansing

        print(dataset.isnull().values.any()) #check whether there is any null value or not
```

```
False
```

Output yang dhasilkan adalah **False**, berarti pada dataset supermarket sales tidak terdapat missing values.

Informasi mengenai dataset adalah sebagai berikut

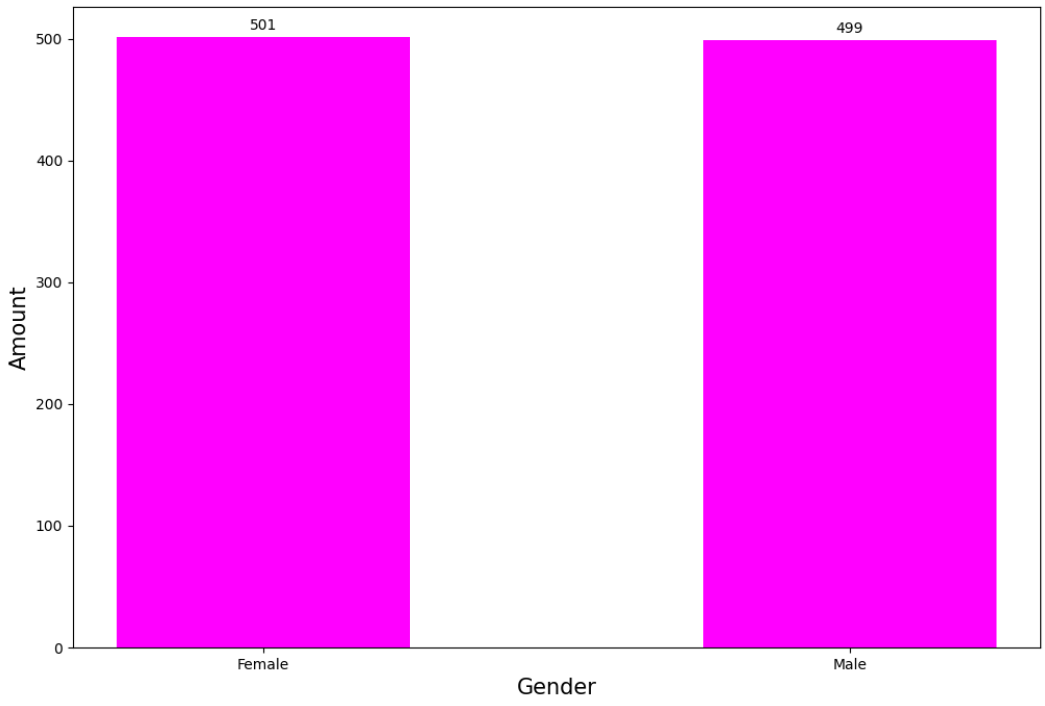
In [5]:
```
print(dataset.info()) #check type
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 17 columns):
 #   Column                  Non-Null Count  Dtype  
---  ------                  --------------  -----  
 0   Invoice ID              1000 non-null   object 
 1   Branch                  1000 non-null   object 
 2   City                    1000 non-null   object 
 3   Customer type           1000 non-null   object 
 4   Gender                  1000 non-null   object 
 5   Product line            1000 non-null   object 
 6   Unit price              1000 non-null   float64
 7   Quantity                1000 non-null   int64  
 8   Tax 5%                  1000 non-null   float64
 9   Total                   1000 non-null   float64
 10  Date                    1000 non-null   object 
 11  Time                    1000 non-null   object 
 12  Payment                 1000 non-null   object 
 13  cogs                    1000 non-null   float64
 14  gross margin percentage  1000 non-null   float64
 15  gross income            1000 non-null   float64
 16  Rating                  1000 non-null   float64
dtypes: float64(7), int64(1), object(9)
memory usage: 132.9+ KB
None
```

Visualisasi

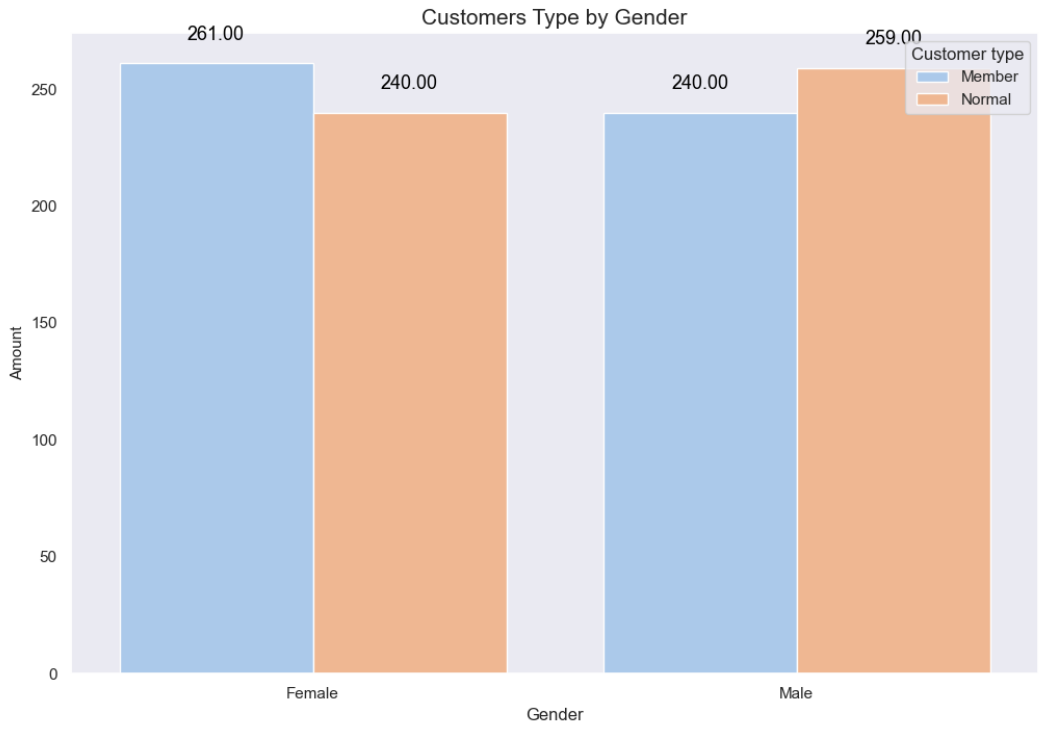
Pertama kita lihat jumlah customer dari supermarket berdasarkan gendernya

In [10]:
```
#Plotting Gender
gender_amount= dataset['Gender'].value_counts()
gender=dataset['Gender'].unique()
fig, ax = plt.subplots(figsize=(12,8))
rects1 = ax.bar(gender,gender_amount,width=0.5,label=gender,color='magenta')
plt.title('Customers by Gender', loc='center', pad=40, fontsize=20, color='black')
plt.ylabel('Amount', fontsize=15)
plt.xlabel('Gender', fontsize=15)
plt.xticks(rotation = 0)
def autolabel(rects):
    """Attach a text label above each bar in *rects*, displaying its height."""
    for rect in rects:
        height = rect.get_height()
        ax.annotate('{}'.format(height),
                    xy=(rect.get_x() + rect.get_width() / 2, height),
                    xytext=(0, 3),  # 3 points vertical offset
                    textcoords="offset points",
                    ha='center', va='bottom')
plt.show()
```



Dari 1000 customer, jumlah pelanggan wanita sebanyak 501 pelanggan dan pria sebanyak 499 pelanggan. Berdasarkan plot ini dapat diketahui bahwa pelanggan dari supermarket cenderung rata antara wanita dan pria.

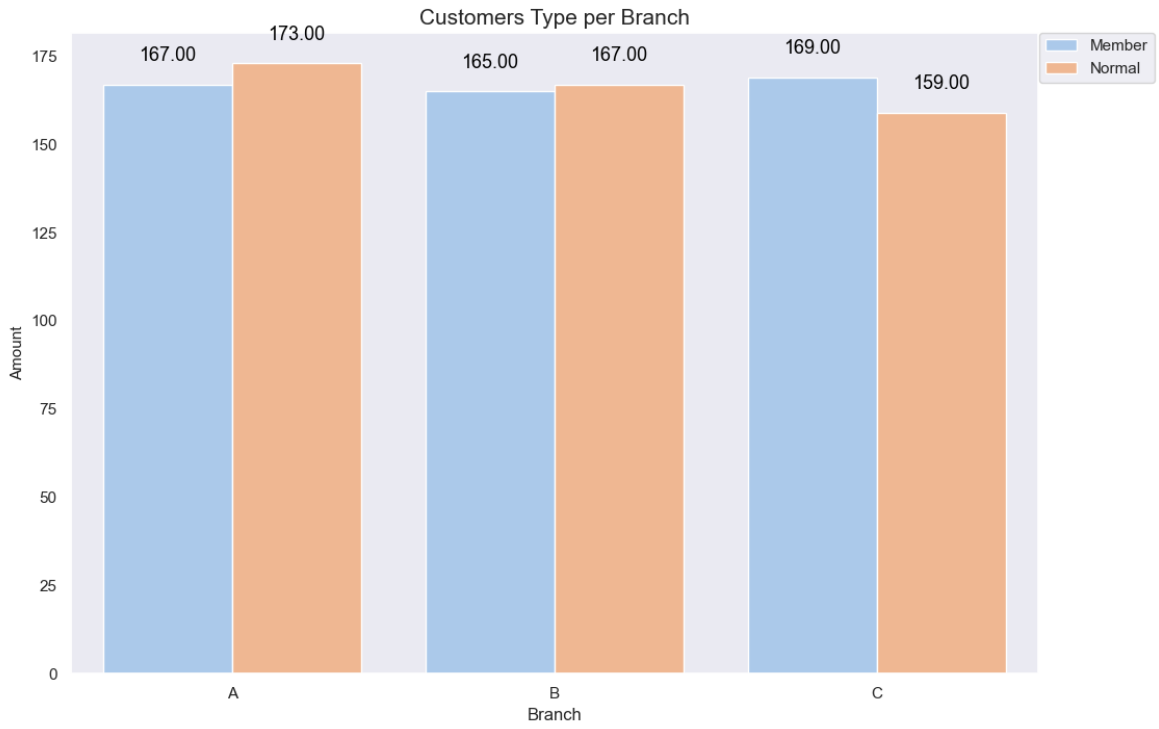
In [12]:
```
#Plotting Cust Type per Gender
groupby_gender = dataset.groupby(['Gender','Customer type'], as_index=False).size()
sns.set({'figure.figsize':(12,8)})
sns.set_theme(style='dark')
rects1 = sns.barplot(x='Gender',y='size',hue='Customer type',data=groupby_gender,palette='pastel')
rects1.set_title('Customers Type by Gender',fontsize=15)
rects1.set_ylabel('Amount',fontsize=11)
for p in rects1.patches:
    rects1.annotate("%.2f" % p.get_height(), (p.get_x() + p.get_width() / 2., p.get_height()),
                    ha='center', va='center', fontsize=13, color='black', rotation=0, xytext=(0, 20),
                    textcoords='offset points')
plt.legend(bbox_to_anchor=(1.12, 1),borderaxespad=0)
plt.show()
```



Dari 501 pelanggan wanita, terdapat 261 pelanggan yang merupakan member dari supermarket dan 240 pelanggan bukan member, sedangkan pelanggan pria lebih banyak yang merupakan bukan member yaitu sebanyak 259 pelanggan dari total 499 pelanggan pria. Jika supermarket ingin menambah jumlah member, salah satu cara yang mungkin bisa dilakukan adalah dengan memberikan promo yang cukup menguntungkan bagi pelanggan yang merupakan member supermarket dibandingkan dengan pelanggan normal.

Sebaran tipe pelanggan berdasarkan cabang supermarket adalah sebagai berikut

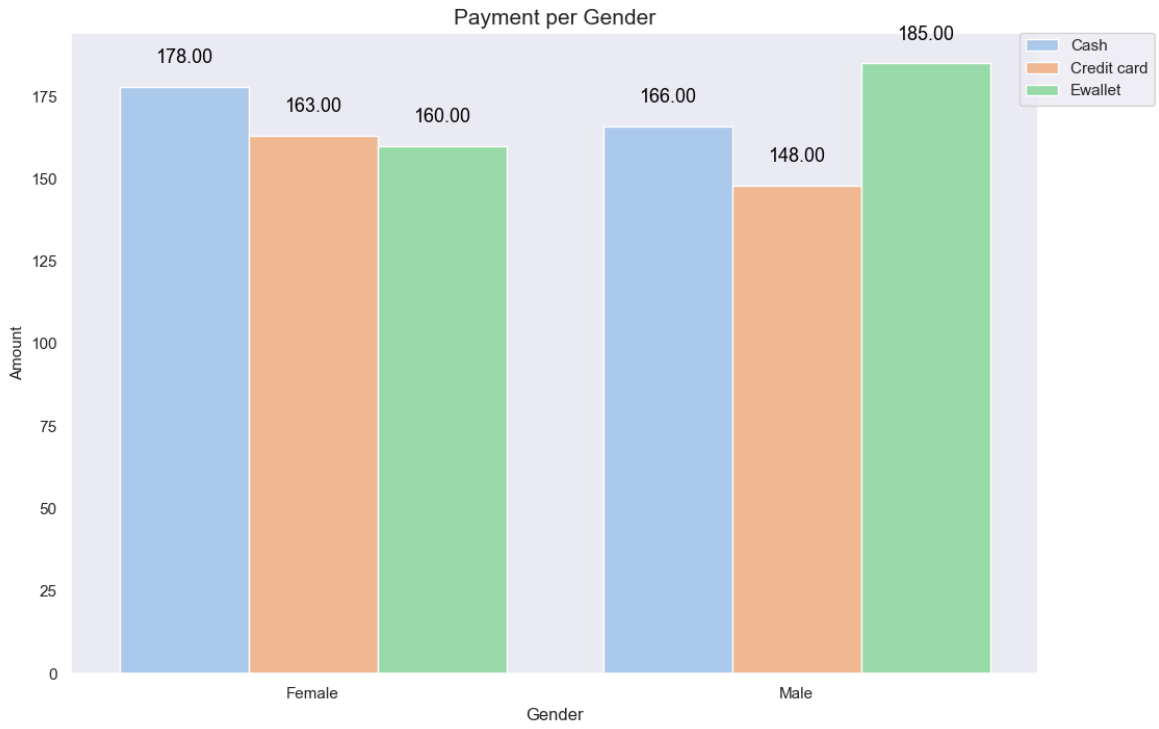
In [5]:
```
#Plotting Customer type per branch
groupby_payment = dataset.groupby(['Branch','Customer type'], as_index=False).size()
plt.figure(figsize=(12,8))
sns.set_theme(style='dark')
rects1 = sns.barplot(x='Branch',y='size',hue='Customer type',data=groupby_payment,palette='pastel')
plt.title('Customers Type per Branch',fontsize=15)
plt.ylabel('Amount',fontsize=11)
for p in rects1.patches:
    rects1.annotate("%.2f" % p.get_height(), (p.get_x() + p.get_width() / 2., p.get_height()),
                    ha='center', va='center', fontsize=13, color='black', rotation=0, xytext=(0, 20),
                    textcoords='offset points')
plt.legend(bbox_to_anchor=(1.12, 1),borderaxespad=0)
plt.show()
```



Di cabang A dan B lebih banyak pelanggan yang merupakan tipe pelanggan normal dengan jumlah yang hampir sama dengan pelanggan member. Berbanding terbalik pada cabang C, dimana lebih banyak pelanggan member dibandingkan dengan pelanggan normal dengan perbedaan yang cukup besar dibanding cabang lainnya.

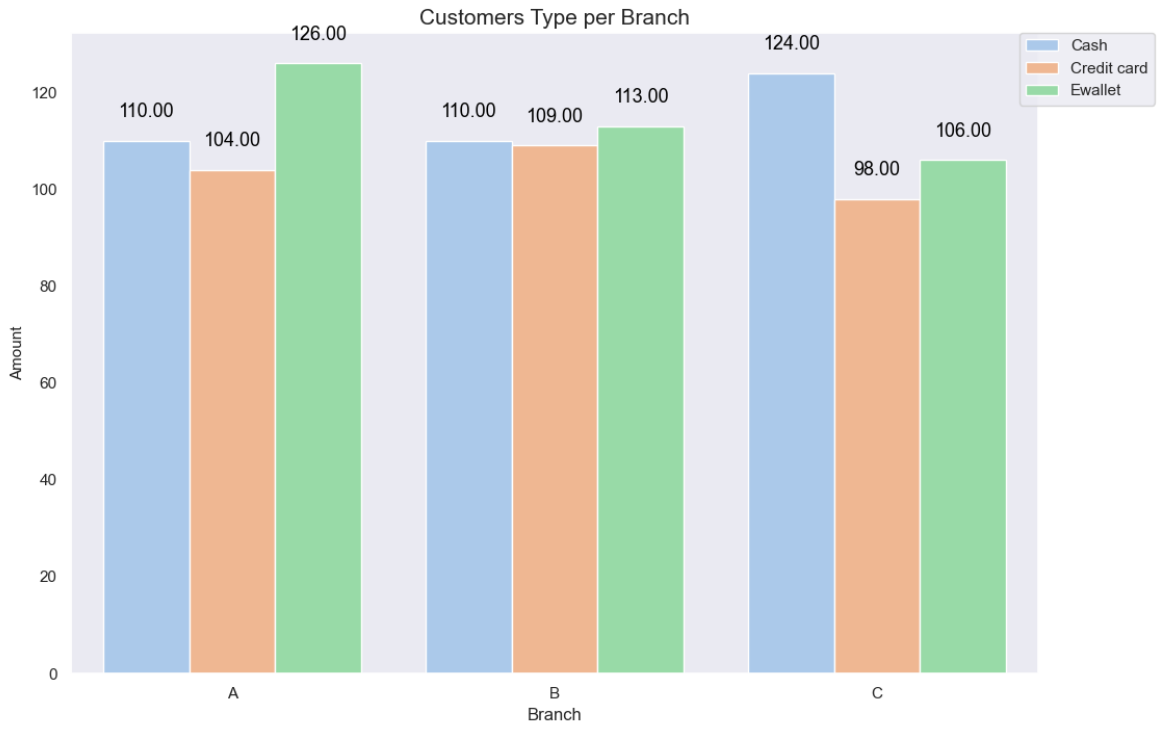
Kemudian, kita lakukan analisis terkait sebaran metode pembayaran berdasarkan gender dan cabang

In [14]:
```
#Plotting Payment by Gender
groupby_paymentg = dataset.groupby(['Gender','Payment'], as_index=False).size()
plt.figure(figsize=(12,8))
sns.set_theme(style='dark')
rects1 = sns.barplot(x='Gender',y='size',hue='Payment',data=groupby_paymentg,palette='pastel')
plt.title('Payment per Gender',fontsize=15)
plt.ylabel('Amount',fontsize=11)
for p in rects1.patches:
    rects1.annotate("%.2f" % p.get_height(), (p.get_x() + p.get_width() / 2., p.get_height()),
                ha='center', va='center', fontsize=13, color='black', rotation=0, xytext=(0, 20),
                textcoords='offset points')
plt.legend(bbox_to_anchor=(1.12, 1),borderaxespad=0)
# plt.show()
```



Terdapat perbedaan behavior cara pembayaran antara pelanggan pria dan wanita, dimana pelanggan wanita lebih memilih membayar belanjaan mereka dengan uang cash sedangkan pelanggan pria lebih memilih untuk membayar dengan menggunakan *e-wallet*.

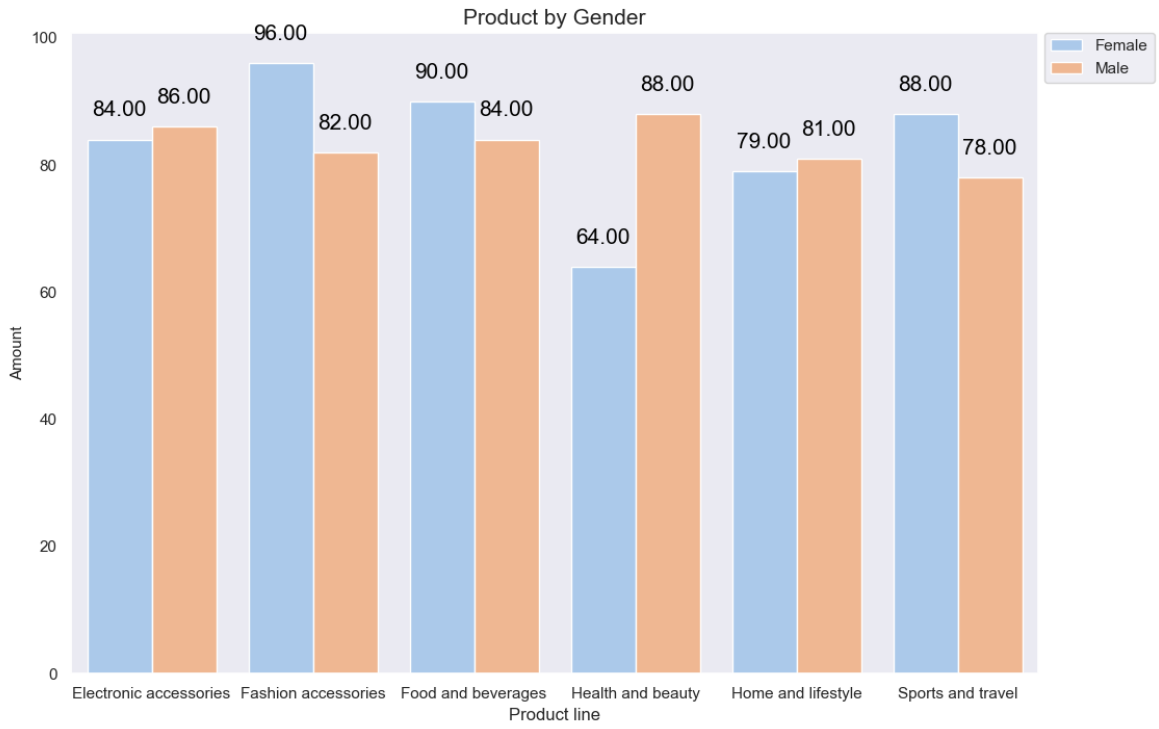
In [6]:
```
#Plotting Payment per branch
groupby_payment = dataset.groupby(['Branch','Payment'], as_index=False).size()
plt.figure(figsize=(12,8))
sns.set_theme(style='dark')
rects1 = sns.barplot(x='Branch',y='size',hue='Payment',data=groupby_payment,palette='pastel')
plt.title('Customers Type per Branch',fontsize=15)
plt.ylabel('Amount',fontsize=11)
for p in rects1.patches:
    rects1.annotate("%.2f" % p.get_height(), (p.get_x() + p.get_width() / 2., p.get_height()),
                ha='center', va='center', fontsize=13, color='black', rotation=0, xytext=(0, 20),
                textcoords='offset points')
plt.legend(bbox_to_anchor=(1.12, 1),borderaxespad=0)
plt.show()
```



Pelanggan di cabang A lebih banyak yang memilih untuk membayar menggunakan *e-wallet* dengan selisih yang cukup jauh dengan metode pembayaran lainnya. Di cabang B, secara umum sebaran pelanggan yang membayar dengan metode pembayaran uang tunai, *credit card*, dan *e-wallet* hampir sama, dengan selisih yang sedikit metode pembayaran *e-wallet* lebih banyak digunakan pada cabang ini. Sedangkan, pada cabang C sebaran penggunaan metode pembayaran cukup jauh, di cabang C metode pembayaran dominan adalah uang tunai.

Selanjutnya, dilakukan analisis terkait penjualan produk

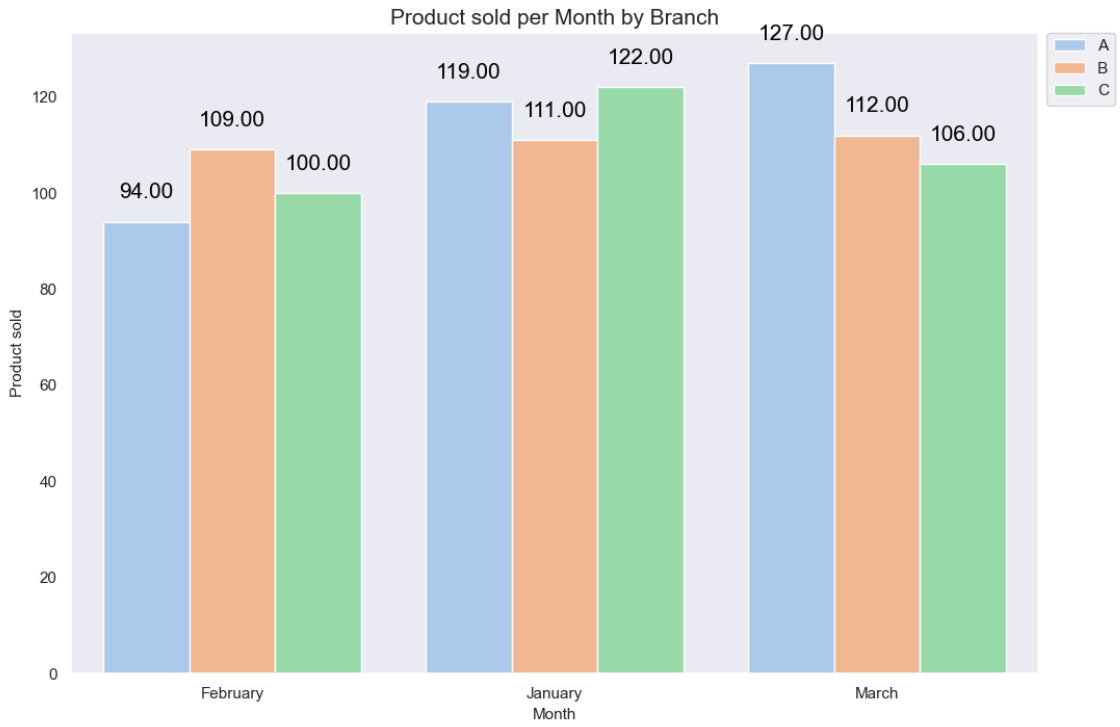
In [16]:
```
#Plotting product line by gender
groupby_productg = dataset.groupby(['Product line','Gender'], as_index=False).size()
plt.figure(figsize=(12,8))
sns.set_theme(style='dark')
rects1 = sns.barplot(x='Product line',y='size',hue='Gender',data=groupby_productg,palette='pastel')
plt.title('Product by Gender',fontsize=15)
plt.ylabel('Amount',fontsize=11)
for p in rects1.patches:
    rects1.annotate("%.2f" % p.get_height(), (p.get_x() + p.get_width() / 2., p.get_height()),
                    ha='center', va='center', fontsize=15, color='black', rotation=0, xytext=(0, 20),
                    textcoords='offset points')
plt.legend(bbox_to_anchor=(1.12, 1),borderaxespad=0)
plt.show()
```



Pelanggan wanita paling banyak membeli produk *fashion accessories* sedangkan pelanggan pria paling banyak membeli produk *health and beauty*.

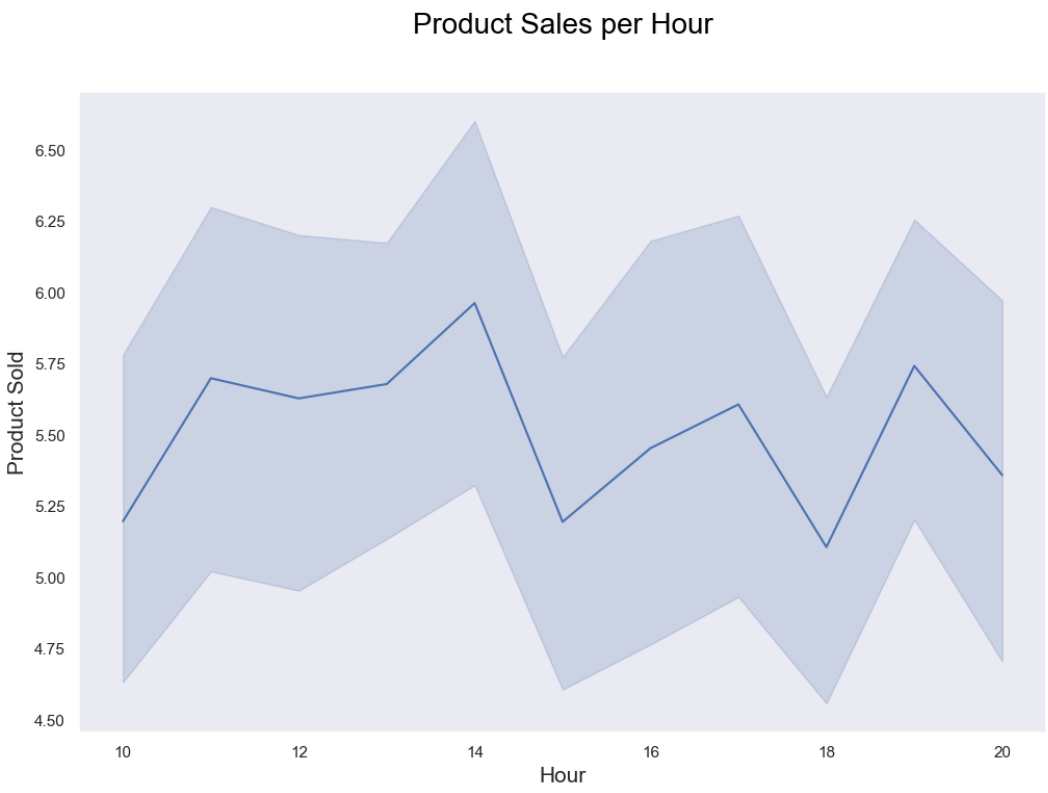
In [9]:
```
dataset['Month'] = dataset['Date'].apply(lambda x:datetime.datetime.strptime(x, "%m/%d/%Y").strftime("%m")) #Creating Month Colu
dataset['month_name'] = dataset['Month'].apply(lambda x:datetime.datetime.strptime(x, "%m").strftime("%B"))
```

In [18]:
```
#Plotting Quantity per Month by Branch
q_month=dataset.groupby(['month_name','Branch'],as_index=False)['Quantity'].size()
plt.figure(figsize=(12,8))
sns.set_theme(style='dark')
rects1 = sns.barplot(x='month_name',y='size',hue='Branch',data=q_month,palette='pastel')
plt.title('Product sold per Month by Branch',fontsize=15)
plt.ylabel('Product sold',fontsize=11)
plt.xlabel('Month',fontsize=11)
for p in rects1.patches:
    rects1.annotate("%.2f" % p.get_height(), (p.get_x() + p.get_width() / 2., p.get_height()),
                    ha='center', va='center', fontsize=15, color='black', rotation=0, xytext=(0, 20),
                    textcoords='offset points')
plt.legend(bbox_to_anchor=(1.08, 1),borderaxespad=0)
plt.show()
```



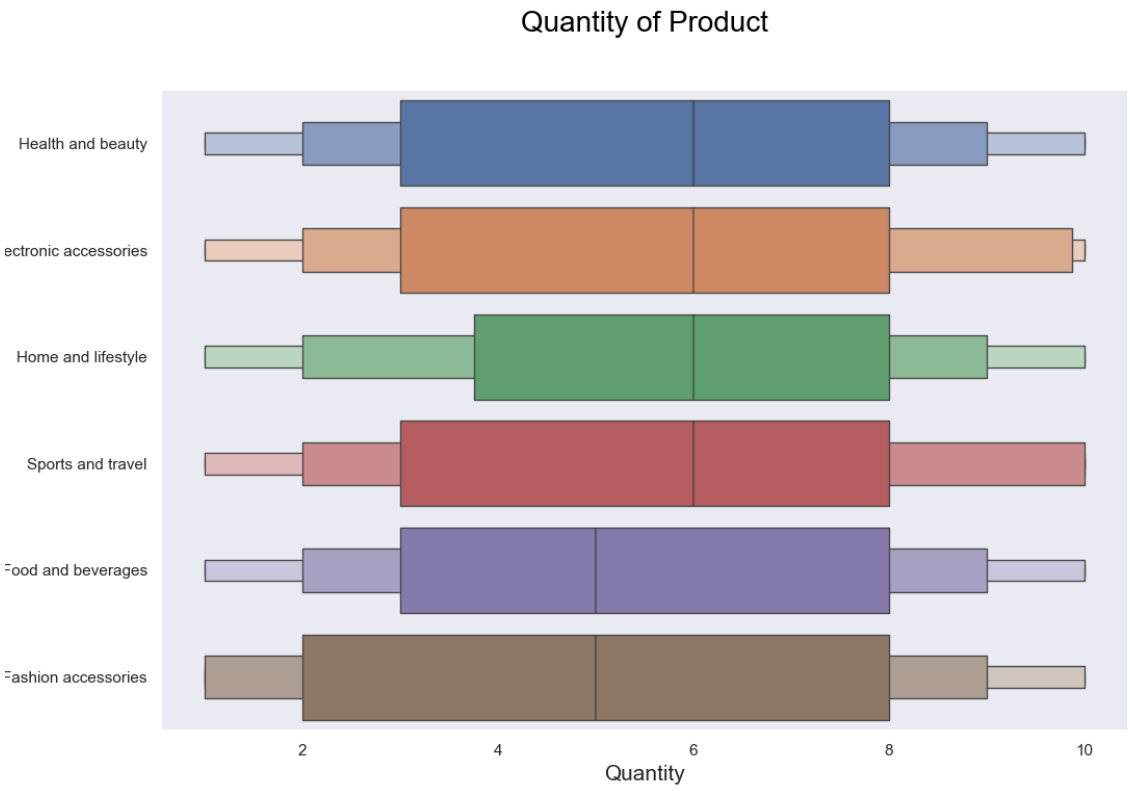
Penjualan produk di cabang A dan C mengalami penurunan cukup tajam pada bulan Februari dan kembali meningkat pada bulan Maret. Sedangkan di cabang B, walaupun terdapat penurunan jumlah produk terjual yang sedikit pada bulan Februari, akan tetapi secara tren penjualan produk di cabang B cenderung konstan.

In [7]:
```
#Plotting Quantity by Hour
dataset['Time']=pd.to_datetime(dataset['Time'])
dataset['Hour']=(dataset['Time']).dt.hour
plt.figure(figsize=(12,8))
sns.set_theme(style='dark')
ax = sns.lineplot(x='Hour',y='Quantity',data=dataset)
ax.set_title('Product Sales per Hour', loc='center', pad=40, fontsize=20, color='black')
ax.set_ylabel('Product Sold', fontsize=15)
ax.set_xlabel('Hour', fontsize=15)
plt.show()
```



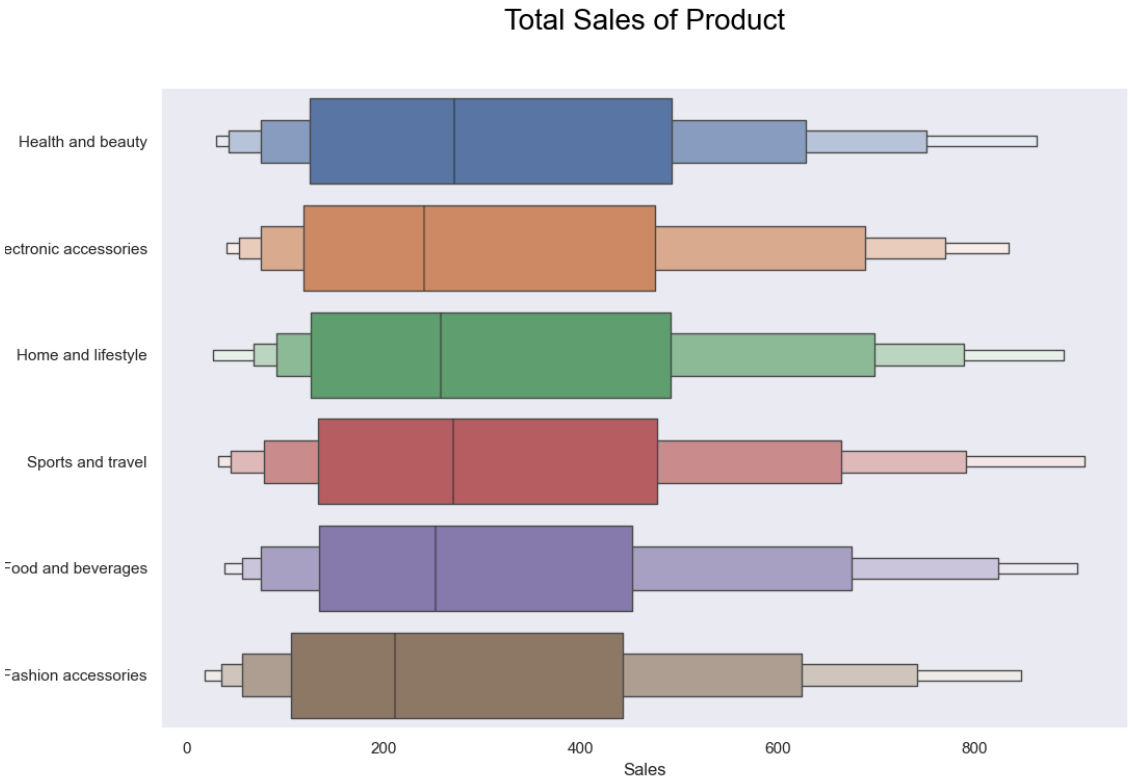
Puncak pembelian terjadi pada pukul 14.00 sehingga dengan mengetahui hal ini supermarket dapat mempersiapkan karyawan yang standby ketika *peak hour*.

In [21]:
```
#Plotting Quantity of Product
plt.figure(figsize=(12,8))
sns.set_theme(style='dark')
ax = sns.boxenplot(x='Quantity',y='Product line',data=dataset)
ax.set_title('Quantity of Product',loc='center', pad=40, fontsize=20, color='black')
ax.set_ylabel('Product', fontsize=15)
ax.set_xlabel('Quantity', fontsize=15)
plt.show()
```



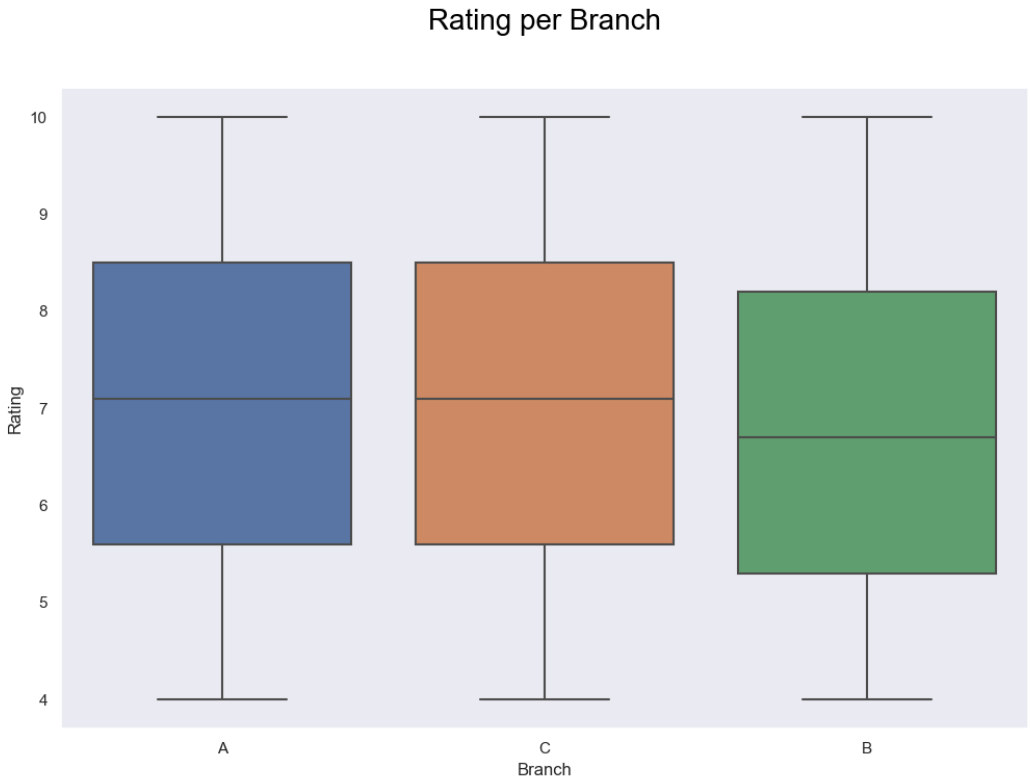
Berdasarkan box plot di atas kita dapat mengetahui kuartil bawah, atas, dan median. Kuartil bawah dan kuartil atas pada produk *home and lifestyle* berturut – turut adalah 3 dan 6 dan pada *fashion accessories* adalah 2 dan 8, sedangkan produk yang lain adalah 3 dan 8 produk. Nilai tengah tertinggi pada pembelian produk adalah 6 buah per produk.

In [20]:
```
#PLotting Total per Product Line
plt.figure(figsize=(12,8))
sns.set_theme(style='dark')
ax = sns.boxenplot(x='Total',y='Product line',data=dataset, showfliers=False)
ax.set_title('Total Sales of Product',loc='center', pad=40, fontsize=20, color='black')
ax.set_ylabel('Product', fontsize=12)
ax.set_xlabel('Sales', fontsize=12)
plt.show()
```



Dari jumlah produk yang terjual diperoleh pendapatan pada masing – masing jenis produk dimana nilai tengah produk *health and beauty* menghasilkan pendapatan yang lebih besar dibandingkan produk lainnya.

In [8]:
```
#Plotting Rating per Branch
plt.figure(figsize=(12,8))
sns.set_theme(style='dark')
ax = sns.boxplot(x='Branch',y='Rating',data=dataset)
ax.set_title('Rating per Branch',loc='center', pad=40, fontsize=20, color='black')
ax.set_ylabel('Rating', fontsize=12)
ax.set_xlabel('Branch', fontsize=12)
plt.show()
```



Nilai tengah penilaian dari pelanggan kepada pelayanan supermarket pada cabang B lebih rendah diantara yang lain. Dari data ini dapat dilakukan peningkatan pelayanan pada cabang B agar penilaian dari pelanggan lebih baik lagi.