Análise Exploratória - Qualidade de Sono

Este documento apresenta a análise exploratória criada para identificar padrões e tendências relacionadas ao sono de uma determinada população utilizando o Python e suas bibliotecas, como Pandas, Matplotlib e Seaborn.

Dados disponíveis em: https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset

1º passo - Instalar e Importar Pacotes e Banco de Dados

In [92]:
```
import pandas as pd # manipulação de dados em formato de dataframe
import numpy as np # operações matemáticas
import seaborn as sns # visualização gráfica
import matplotlib.pyplot as plt # visualização gráfica
import pingouin as pg # outro modo para obtenção de matrizes de correlações
from scipy.stats import boxcox # transformação de Box-Cox
from matplotlib import font_manager
```

In [10]:
```
df = pd.read_excel("Qualidade de Sono.xlsx")
```

2º passo - Organizacao dos Dados

In
```
df = df.rename(columns={'Person ID': 'Pessoa_ID',
                        'Gender': 'Sexo',
                        'Age': 'Idade',
                        'Occupation': 'Cargo',
                        'Sleep Duration': 'Duracao_Sono_h',
                        'Quality of Sleep': 'Qualidade_Sono',
                        'Physical Activity Level': 'Nível_Atividade_Fisica_h',
                        'Stress Level': 'Nivel_Estresse',
                        'BMI Category': 'Categoria_IMC',
                        'Blood Pressure': 'Pressao_Arterial',
                        'Heart Rate': 'Frequencia_Cardiaca',
                        'Daily Steps': 'Passos_Diarios',
                        'Sleep Disorder': 'Disturbio_do_Sono'})

print(df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 374 entries, 0 to 373
Data columns (total 13 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   Pessoa_ID                 374 non-null    int64  
 1   Sexo                      374 non-null    object 
 2   Idade                     374 non-null    int64  
 3   Cargo                     374 non-null    object 
 4   Duracao_Sono_h            374 non-null    float64
 5   Qualidade_Sono            374 non-null    int64  
 6   Nível_Atividade_Fisica_h  374 non-null    int64  
 7   Nivel_Estresse            374 non-null    int64  
 8   Categoria_IMC             374 non-null    object 
 9   Pressao_Arterial          374 non-null    object 
 10  Frequencia_Cardiaca       374 non-null    int64  
 11  Passos_Diarios            374 non-null    int64  
 12  Disturbio_do_Sono         155 non-null    object 
dtypes: float64(1), int64(7), object(5)
memory usage: 38.1+ KB
None
```

Aqui percebemos um erro na coluna **Disturbio_do_Sono**, onde ela apresenta uma ausência de dados em relacão as outras colunas. É preciso preencher as colunas vazias para evitar futuros problemas e erros de análise. Além disso, a coluna **Pessoa_ID** não será útil para nossa análise.Para isso, são feitos os seguintes processos:

In [14]:
```
## Remover Coluna Pessoa_ID

df = df.drop('Pessoa_ID', axis=1)

## Preenchendo células vazias

df['Disturbio_do_Sono'].fillna('No Sleep Disorder',inplace=True)

print(df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 374 entries, 0 to 373
Data columns (total 12 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   Sexo                      374 non-null    object 
 1   Idade                     374 non-null    int64  
 2   Cargo                     374 non-null    object 
 3   Duracao_Sono_h            374 non-null    float64
 4   Qualidade_Sono            374 non-null    int64  
 5   Nível_Atividade_Fisica_h  374 non-null    int64  
 6   Nivel_Estresse            374 non-null    int64  
 7   Categoria_IMC             374 non-null    object 
 8   Pressao_Arterial          374 non-null    object 
 9   Frequencia_Cardiaca       374 non-null    int64  
 10  Passos_Diarios            374 non-null    int64  
 11  Disturbio_do_Sono         374 non-null    object 
dtypes: float64(1), int64(6), object(5)
memory usage: 35.2+ KB
None
```

Como observado, as colunas que estavam sem valores foram preenchidas e a coluna Pessoa_ID foi removida. Agora vamos analisar as estatísitcas das variáveis:

In [16]:
```
tab_desc = df.describe()

print(tab_desc)
```

```
             Idade  Duracao_Sono_h   Qualidade_Sono    Nível_Atividade_Fisica_h  \
count   374.000000      374.000000       374.000000                374.000000
mean     42.184492        7.132086         7.312834                 59.171123
std       8.673133        0.795657         1.196956                 20.830804
min      27.000000        5.800000         4.000000                 30.000000
25%      35.250000        6.400000         6.000000                 45.000000
50%      43.000000        7.200000         7.000000                 60.000000
75%      50.000000        7.800000         8.000000                 75.000000
max      59.000000        8.500000         9.000000                 90.000000

       Nivel_Estresse   Frequencia_Cardiaca   Passos_Diarios
count      374.000000            374.000000       374.000000
mean         5.385027             70.165775      6816.844920
std          1.774526              4.135676      1617.915679
min          3.000000             65.000000      3000.000000
25%          4.000000             68.000000      5600.000000
50%          5.000000             70.000000      7000.000000
75%          7.000000             72.000000      8000.000000
max          8.000000             86.000000     10000.000000
```

Separando as profissões por área

In [19]:
```
area = {
    'Nurse': 'Saúde',
    'Doctor': 'Saúde',
    'Engineer': 'Tecnologia',
    'Software Engineer': 'Tecnologia',
    'Scientist': 'Tecnologia',
    'Lawyer': 'Direito',
    'Teacher': 'Educação',
    'Salesperson': 'Comercial',
    'Sales Representative': 'Comercial',
    'Accountant': 'Comercial',
    'Manager': 'Comercial'
}

df['Area'] = df['Cargo'].map(area)
```

Criando faixas Etárias

In [21]:
```
limites_faixas_etarias = [26, 35, 45, 55, 60]
faixa_etaria = ['27-35', '36-45', '46-55', '55+']

df['faixa_etaria'] = pd.cut(df['Idade'], bins = limites_faixas_etarias, labels=
```

Convertendo Nível de Atividade Física para Horas (Está em minutos)

In [29]:
```
df['Nível_Atividade_Fisica_h'] = df['Nível_Atividade_Fisica_h'] /60

df['Nível_Atividade_Fisica_h'] = df['Nível_Atividade_Fisica_h'].round(1)
```

3º passo - Análises Demográficas

In [32]:
```
#%% Sexo

cont_sexo = df['Sexo'].value_counts()

percent_sexo = (cont_sexo / cont_sexo.sum()) * 100

percent_sexo = percent_sexo.round(2)

resumo_sexo = pd.DataFrame({'Contagem' : cont_sexo, '%' : percent_sexo})

print(resumo_sexo)
```

```
        Contagem      %
Sexo
Male          189  50.53
Female        185  49.47
```

A base de dados apresenta uma quantidade equilibrada de homens e mulheres, com uma diferenca de apenas 4 homens a mais que mulheres.

In [35]:
```
#%% Idade

cont_idade = df['faixa_etaria'].value_counts()

percent_idade = (cont_idade / cont_idade.sum()) * 100

percent_idade = percent_idade.round(1)

resumo_idade = pd.DataFrame({'Contagem' : cont_idade, '%' : percent_idade})

print(resumo_idade)
```

```
              Contagem     %
faixa_etaria
36-45              170  45.5
27-35               94  25.1
46-55               77  20.6
55+                 33   8.8
```

A maior parte das pessoas na base de dados tem entre 36 a 45 anos.

In [38]:
```
#%% Peso

cont_peso = df['Categoria_IMC'].value_counts()

percent_peso = (cont_peso / cont_peso.sum()) * 100

percent_peso = percent_peso.round(1)

resumo_peso = pd.DataFrame({'Contagem' : cont_peso, '%' : percent_peso})
```

```
print(resumo_peso)
```

```
                Contagem     %
Categoria_IMC
Normal               216  57.8
Overweight           148  39.6
Obese                 10   2.7
```

A base de dados contém um número significativo de pessoas com peso normal e acima do peso, enquanto a minoria é composta por pessoas obesas.

In [41]:
```
#%% Disturbio de sono

cont_disturbio = df['Disturbio_do_Sono'].value_counts()

percent_disturbio = (cont_disturbio / cont_disturbio.sum ()) * 100

percent_disturbio = percent_disturbio.round(1)

resumo_peso = pd.DataFrame({'Contagem' : cont_disturbio, '%' : percent_disturbio

print(resumo_peso)
```

```
                    Contagem     %
Disturbio_do_Sono
No Sleep Disorder        219  58.6
Sleep Apnea               78  20.9
Insomnia                  77  20.6
```

A maior parte das pessoas na base de dados não apresenta nenhum tipo de distúrbio do sono, e a quantidade de indivíduos que têm distúrbios é equilibrada, com metade sofrendo de apneia do sono e a outra metade de insônia.

In [44]:
```
## Análise das profissões

cont_profissoes = df['Cargo'].value_counts()

print(cont_profissoes)
```

```
Cargo
Nurse                    73
Doctor                   71
Engineer                 63
Lawyer                   47
Teacher                  40
Accountant               37
Salesperson              32
Software Engineer         4
Scientist                 4
Sales Representative      2
Manager                   1
Name: count, dtype: int64
```

In [46]:
```
## Análise das Áreas de profissão

cont_area = df['Area'].value_counts()

print(cont_area)
```

```
Area
Saúde          144
Comercial       72
Tecnologia      71
Direito         47
Educação        40
Name: count, dtype: int64
```

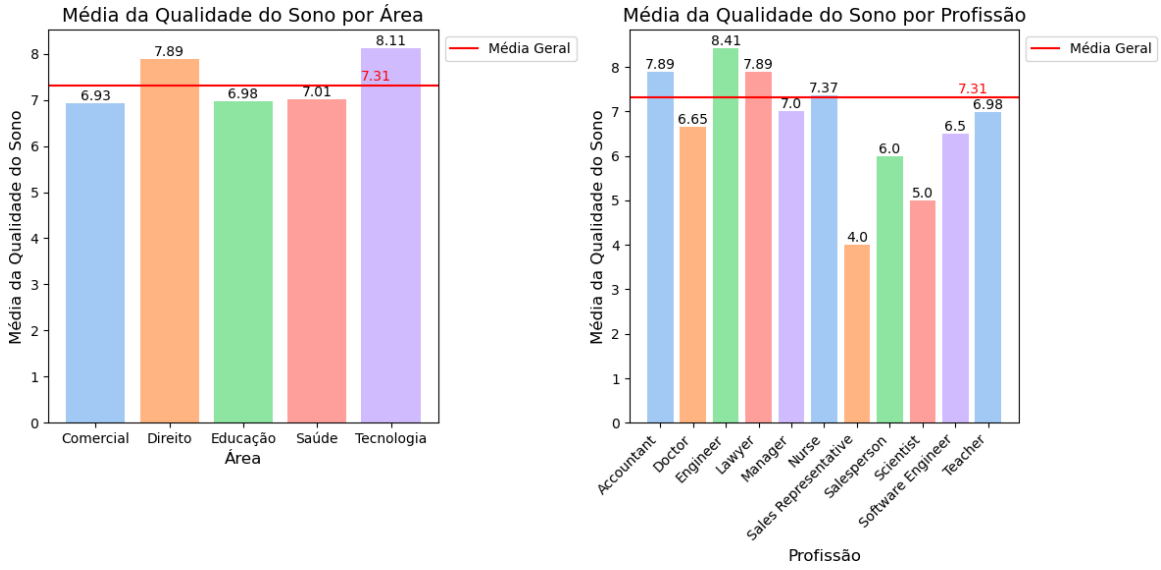
A maior parte das pessoas na base são da área da Saúde, seguidas por pessoas da área Comercial e Tecnologia.

4º passo - Análise da Qualidade do Sono da População

Qualidade de Sono por Área e Profissão

In [51]:
```
media_sono_area = df.groupby('Area')['Qualidade_Sono'].mean().reset_index()
colors = sns.color_palette('pastel')[0:5]

plt.figure(figsize=(12, 6))

# Gráfico para a média por área
plt.subplot(1, 2, 1)
bars_area = plt.bar(media_sono_area['Area'], media_sono_area['Qualidade_Sono'],
plt.title('Média da Qualidade do Sono por Área', fontsize=14)
plt.xlabel('Área', fontsize=12)
plt.ylabel('Média da Qualidade do Sono', fontsize=12)

for bar in bars_area:
    yval = bar.get_height()
    plt.text(bar.get_x() + bar.get_width() / 2, yval, round(yval, 2), ha='center

media_geral_area = df['Qualidade_Sono'].mean()
plt.axhline(y=media_geral_area, color='red', linestyle='-', label='Média Geral')
plt.text(x=len(media_sono_area) - 1, y=media_geral_area + 0.1,
         s=f'{round(media_geral_area, 2)}',
         color='red', ha='right', fontsize=10)

plt.legend(loc='upper left', bbox_to_anchor=(1, 1))

# Gráfico para a média da qualidade do sono por profissão
media_sono_profissao  =  df.groupby('Cargo')['Qualidade_Sono'].mean().reset_index(

plt.subplot(1, 2, 2)  # 1 linha, 2 colunas, 2º gráfico
bars_profissao = plt.bar(media_sono_profissao['Cargo'], media_sono_profissao['Qu
plt.title('Média da Qualidade do Sono por Profissão', fontsize=14)
plt.xlabel('Profissão', fontsize=12)
plt.ylabel('Média da Qualidade do Sono', fontsize=12)
plt.xticks(rotation=45, ha='right')

for bar in bars_profissao:
    yval = bar.get_height()
    plt.text(bar.get_x() + bar.get_width() / 2, yval, round(yval, 2), ha='center

media_geral_profissao = df['Qualidade_Sono'].mean()
plt.axhline(y=media_geral_profissao, color='red', linestyle='-', label='Média Ge
plt.text(x=len(media_sono_profissao) - 1, y=media_geral_profissao + 0.1,
```

```
        s=f'{round(media_geral_profissao, 2)}',
        color='red', ha='right', fontsize=10)

plt.legend(loc='upper left', bbox_to_anchor=(1, 1))

plt.tight_layout()

plt.show()
```



A partir da análise acima, percebemos que pessoas da área da Educação, Saúde e Comercial tem uma qualidade do sono abaixo da média. Percebemos também que Representantes de Vendas tem a pior qualidade de sono (4.0) dentre todas profissões registradas, junto com Cientistas (5.0) e Vendedores (6.0).

Qualidade de Sono relacionado a Distúrbio do Sono

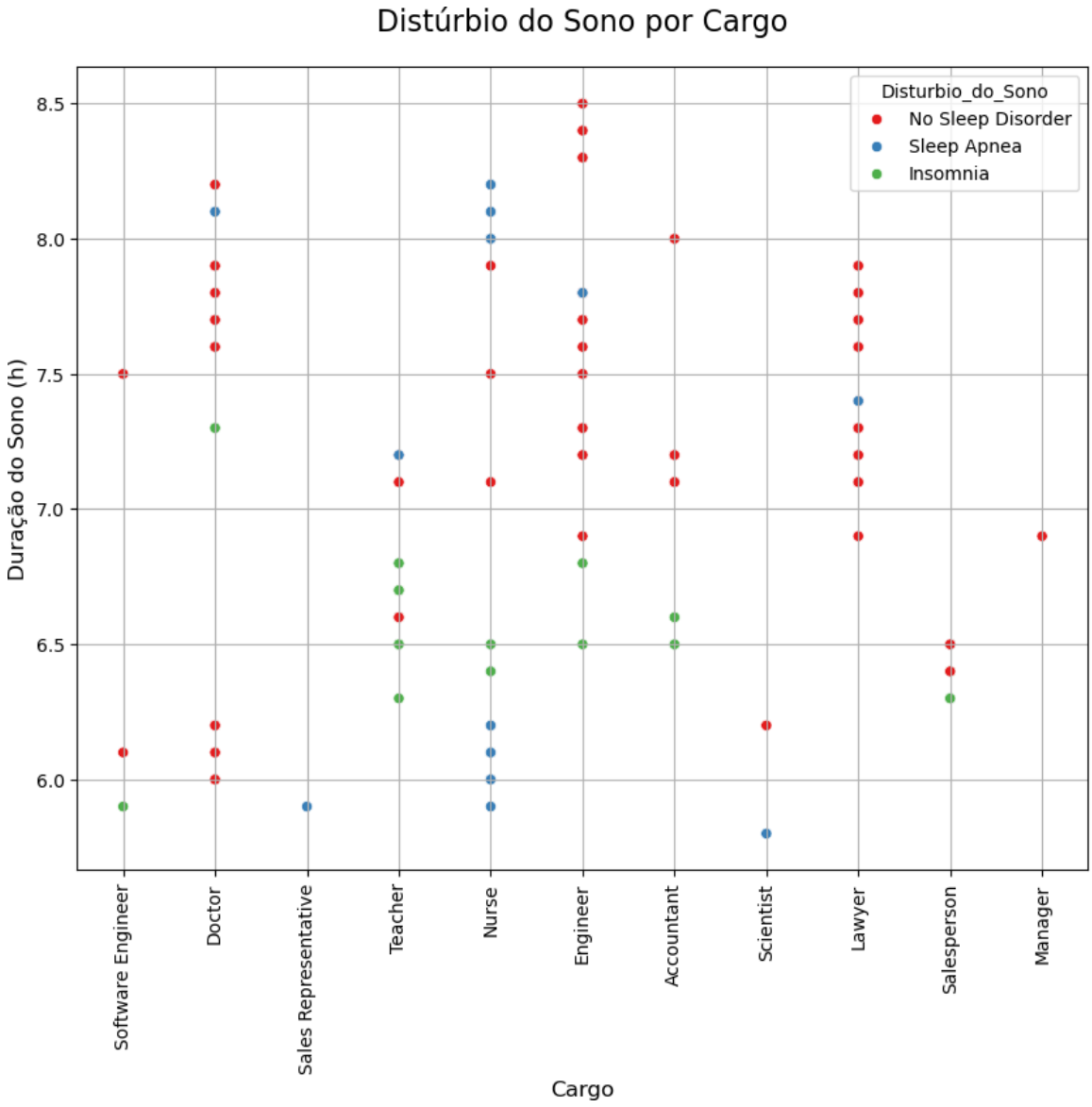
In [55]:
```
media_sono_semdisturb = df[df['Disturbio_do_Sono'].isnull()]['Qualidade_Sono'].m

print('Media sem disturbio: ',  media_sono_semdisturb)

media_sono_comdisturb  =  df.groupby('Disturbio_do_Sono')['Qualidade_Sono'].mean()

print('Media com disturbio: \n',  media_sono_comdisturb)
```

```
Media sem disturbio:  nan
Media com disturbio:
    Disturbio_do_Sono  Qualidade_Sono
0           Insomnia        6.532468
1  No Sleep Disorder        7.625571
2        Sleep Apnea        7.205128
```

Aqui percebemos que a insônia é o distúrbio que mais influencia na qualidade de sono das pessoas.

In [58]:
```
plt.figure(figsize=(10,  8))

sns.scatterplot(data=df, x='Cargo', y='Duracao_Sono_h', hue='Disturbio_do_Sono',

plt.title('Distúrbio do Sono por Cargo', fontsize=16, pad=20,)
```

```
plt.xlabel('Cargo', fontsize=12)
plt.ylabel('Duração do Sono (h)', fontsize=12)
plt.axhline(y=media_sono_semdisturb, color='red', linestyle='--', label=f'Média:

plt.grid(True)

plt.xticks(rotation=90)
plt.show()
```



Percebemos, a partir do gráfico acima, que enfermeiros e professores tem tendência a desenvolver distúrbios do sono em comparacão com outras profissões. Enfermeiros tendem a desenvolver Apneia do Sono, o que explica a qualidade do sono acima da média, visto que este distúrbio não afeta tanto na qualidade do sono de seus portadores, e professores tendem a desenvolver Insônia, o que pode explicar o motivo da qualidade do sono ser abaixo da média .

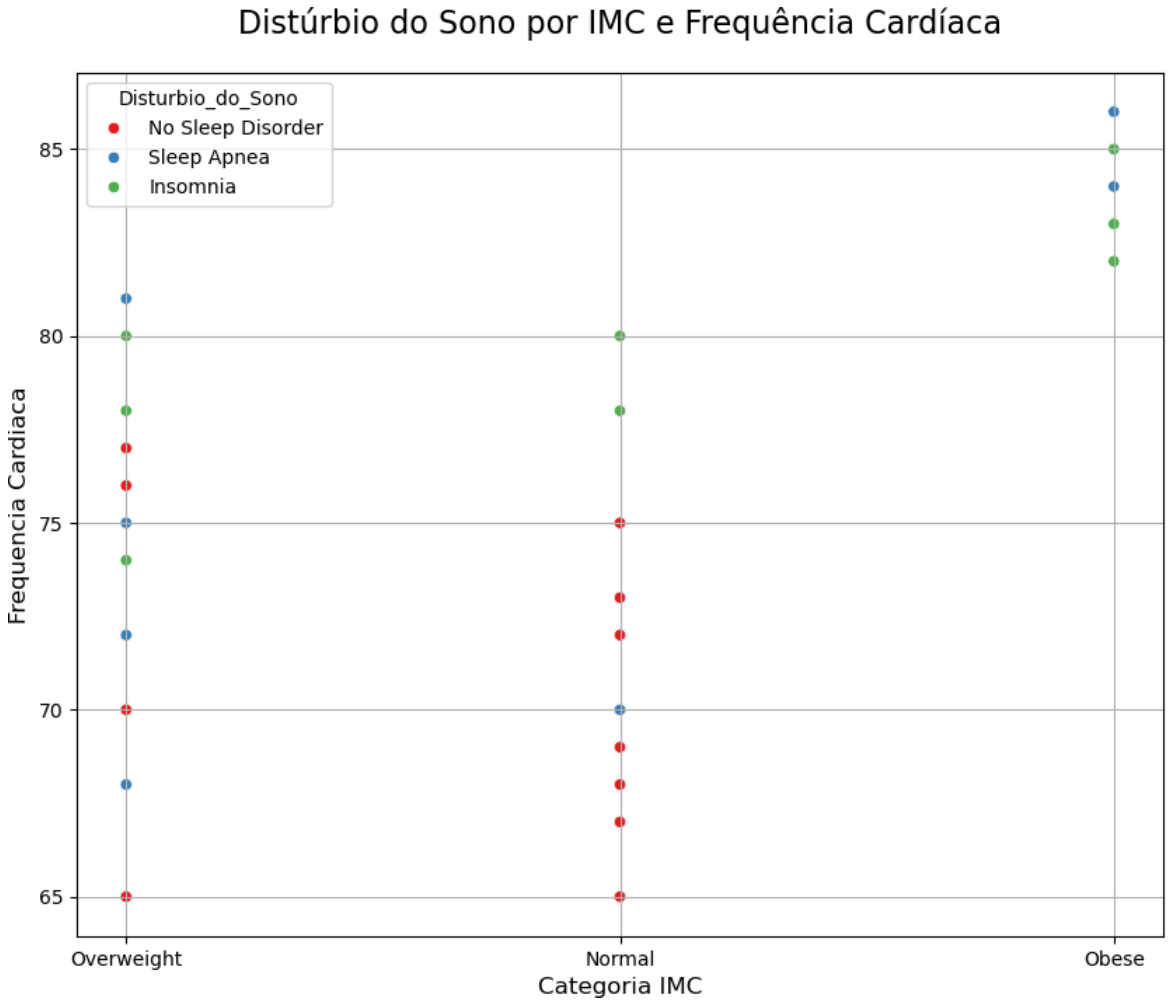
Qualidade de Sono relacionado ao IMC das pessoas

In [73]:
```
media_sono_imc=  df.groupby('Categoria_IMC')['Qualidade_Sono'].mean().reset_index

print('Media com disturbio: \n',  media_sono_imc)
```

```
Media com disturbio:
   Categoria_IMC   Qualidade_Sono
0         Normal         7.638889
1          Obese         6.400000
2     Overweight         6.898649
```

Aqui percebemos que, quanto maior o peso das peessoas, pior a qualidade de sono.

In [88]:
```
plt.figure(figsize=(10, 8))

sns.scatterplot(data=df, x='Categoria_IMC', y='Frequencia_Cardiaca', hue='Distur

plt.title('Distúrbio do Sono por IMC e Frequência Cardíaca', fontsize=16, pad=20
plt.xlabel('Categoria IMC', fontsize=12)
plt.ylabel('Frequencia Cardiaca', fontsize=12)

plt.grid(True)

plt.xticks()
plt.show()
```



Percebemos, a partir do gráfico acima, que o peso e a frequência cardíaca são fatores extremamente relevantes quando relacionados ao distúrbio de sono, visto que quanto maior o peso das pessoas, maior será frequência cardíacia e a tendência a desenvolver distúrbios de sono.