preprocessing

January 25, 2025

1 Limpieza de datos de la encuesta

1.1 Cargamos librerías

[1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import json
import unidecode
import missingno as msno
from scipy.stats import chi2_contingency
import utils
import yaml
```

1.2 Cargamos los datos

- Carga de datos originales
- Cambio de nombre de las variables según diccionario

[2]:
```
__path__ = '/Users/iairlinker/Documents/repos/taller_ahh1/data/raw/'
# Load the encuesta DataFrame
encuesta = pd.read_csv(__path__ + 'encuesta.csv')
encuesta.columns
```

[2]:
```
Index(['Marca temporal', 'Dirección de correo electrónico', 'Nombres',
       'Apellidos', 'Fecha de nacimiento', '¿Qué tanto te gusta estudiar?',
       '¿Cuántas horas te dedicarías en una semana a estudiar para una
asignatura si no tuvieras una evaluación pronto?',
       '¿Cuánto tiempo dedicas en una semana a estudiar para una asignatura en
la que pronto tendrás una evaluación?',
       'Si sientes que estás preparad(a/o) para una evaluación ¿Dedicarías horas
a estudiar de todas formas?',
       'Cuando estudias algo relacionado con Matemática…¿Cuántos ejercicios
resuelves en una sesión de estudio?',
       '¿Cuál o cuáles de los siguientes métodos utilizas para estudiar?',
       '¿En qué horario prefieres estudiar?',
       '¿Qué lugar(es) utilizas para estudiar?',
```

```
        '¿Qué factores consideras que dificultan tus estudios?',
        '¿Sientes que tienes tiempo suficiente para estudiar?',
        '¿Cuántas horas dedicas diariamente a dormir?',
        '¿Cuántas horas dedicas diariamente a descansar?',
        '¿Estudias sol(a/o) o acompañad(a/o)?',
        '¿Si no sabes resolver un problema a quién acudes?',
        'Por favor describe brevemente por qué estás estudiando esta carrera',
        '¿Has tenido la idea de retirarte o cambiarte a otra carrera?',
        'Solo si la respuesta anterior fue 'Sí' indica cuándo has pensado en
cambiarte a otra carrera o retirarte'],
      dtype='object')
```

[3]:
```
encuesta.applymap(lambda x: x.strip() if isinstance(x, str) else x)
encuesta.head().T
```

```
/var/folders/02/wzc8k1dn7bsgd14yfq9md5l40000gn/T/ipykernel_24203/1935490141.py:1
: FutureWarning: DataFrame.applymap has been deprecated. Use DataFrame.map
instead.
  encuesta.applymap(lambda x: x.strip() if isinstance(x, str) else x)
```

[3]:
```
                                                    0  \
Marca temporal                                         
9/12/2024 9:53:35                                      
Dirección de correo electrónico                        
isidorahuaiquilaf402@gmail.com                         
Nombres                                                
Isidora Esperanza                                      
Apellidos                                              
Huaiquilaf Verdugo                                     
Fecha de nacimiento                                    
3/10/2004                                              
¿Qué tanto te gusta estudiar?                          
2                                                      
¿Cuántas horas te dedicarías en una semana a es…       
Entre una y dos horas                                  
¿Cuánto tiempo dedicas en una semana a estudiar…  Estudio hasta que siento que
estoy preparad(a/…                                     
Si sientes que estás preparad(a/o) para una eva…       
Definitivamente sí                                     
Cuando estudias algo relacionado con Matemática…  Todos los necesarios (hasta
que siento que est…                                    
¿Cuál o cuáles de los siguientes métodos utiliz…  Resolver ejercicios, Hacer
resúmenes, Elaborar…                                   
¿En qué horario prefieres estudiar?                 En la mañana (desde que me
levanto hasta medio…                                   
¿Qué lugar(es) utilizas para estudiar?              Una habitación, Una pieza de
estudio, Una bibl…                                     
¿Qué factores consideras que dificultan tus est…       
```

Estrés
¿Sientes que tienes tiempo suficiente para estu…
3
¿Cuántas horas dedicas diariamente a dormir?
6
¿Cuántas horas dedicas diariamente a descansar?
3
¿Estudias sol(a/o) o acompañad(a/o)?
Sol(a/o)
¿Si no sabes resolver un problema a quién acudes?
Profesor(a)
Por favor describe brevemente por qué estás est… Quiero saber constituir una empresa y me gusta…
¿Has tenido la idea de retirarte o cambiarte a …
Sí
Solo si la respuesta anterior fue 'Sí' indica c… Momentos de muchas pruebas y evaluaciones junt…

1 \
Marca temporal
9/12/2024 10:10:46
Dirección de correo electrónico
valentina.adm.ite@gmail.com
Nombres
Valentina del carmen
Apellidos
Silva ite
Fecha de nacimiento
13/07/2005
¿Qué tanto te gusta estudiar?
3
¿Cuántas horas te dedicarías en una semana a es…
Entre una y dos horas
¿Cuánto tiempo dedicas en una semana a estudiar…
Entre una y dos horas
Si sientes que estás preparad(a/o) para una eva…
Tal vez
Cuando estudias algo relacionado con Matemática… Ninguno.
Solo reviso el material
¿Cuál o cuáles de los siguientes métodos utiliz… Resolver ejercicios, Explicar los temas en voz…
¿En qué horario prefieres estudiar? En la noche (desde el atardecer hasta que me v…
¿Qué lugar(es) utilizas para estudiar? Una habitación, Una biblioteca
¿Qué factores consideras que dificultan tus est…
Falta de tiempo

¿Sientes que tienes tiempo suficiente para estu…
1
¿Cuántas horas dedicas diariamente a dormir?
5 horas
¿Cuántas horas dedicas diariamente a descansar?
Solo cuando duermo
¿Estudias sol(a/o) o acompañad(a/o)?
Sol(a/o)
¿Si no sabes resolver un problema a quién acudes? A un amigo o amiga cercana
Por favor describe brevemente por qué estás est… Por qué me gusta el marketing y la relaciones …
¿Has tenido la idea de retirarte o cambiarte a …
No
Solo si la respuesta anterior fue 'Sí' indica c…
NaN

2 \
Marca temporal
9/12/2024 10:12:27
Dirección de correo electrónico
aljaque@alumnos.uahurtado.cl
Nombres
Alexis Alexander
Apellidos
Jaque Cardenas
Fecha de nacimiento
30/06/2006
¿Qué tanto te gusta estudiar?
4
¿Cuántas horas te dedicarías en una semana a es… Estudiaría hasta que sienta que estoy preparad…
¿Cuánto tiempo dedicas en una semana a estudiar…
Entre 4 y 7 horas
Si sientes que estás preparad(a/o) para una eva…
Definitivamente sí
Cuando estudias algo relacionado con Matemática… Todos los necesarios (hasta que siento que est…
¿Cuál o cuáles de los siguientes métodos utiliz… Resolver ejercicios, Hacer resúmenes, Elaborar…
¿En qué horario prefieres estudiar? Me es indiferente (cualquier horario es bueno …
¿Qué lugar(es) utilizas para estudiar? Una habitación, Una pieza de estudio, Una bibl…
¿Qué factores consideras que dificultan tus est… Puede ser falta de recursos y tiempo, ya que t…
¿Sientes que tienes tiempo suficiente para estu…

3
¿Cuántas horas dedicas diariamente a dormir?
10 u 8
¿Cuántas horas dedicas diariamente a descansar?
2 horas
¿Estudias sol(a/o) o acompañad(a/o)?
Sol(a/o)
¿Si no sabes resolver un problema a quién acudes?
Profesor(a)
Por favor describe brevemente por qué estás est… Porque me llama la atención el comercio
¿Has tenido la idea de retirarte o cambiarte a …
No
Solo si la respuesta anterior fue 'Sí' indica c…
NaN

3 \
Marca temporal
9/12/2024 10:16:27
Dirección de correo electrónico
Naty20godoydu@gmail.com
Nombres
Natalia
Apellidos
Godoy
Fecha de nacimiento
11/11/2005
¿Qué tanto te gusta estudiar?
3
¿Cuántas horas te dedicarías en una semana a es…
Entre 2 y 4 horas
¿Cuánto tiempo dedicas en una semana a estudiar…
Entre 4 y 7 horas
Si sientes que estás preparad(a/o) para una eva…
Definitivamente sí
Cuando estudias algo relacionado con Matemática…
Entre 1 y 5
¿Cuál o cuáles de los siguientes métodos utiliz… Hacer resúmenes, Leer y subrayar, Ver vídeos e…
¿En qué horario prefieres estudiar? En la noche (desde el atardecer hasta que me v…
¿Qué lugar(es) utilizas para estudiar? Una habitación, El patio, Una pieza de estudio
¿Qué factores consideras que dificultan tus est…
Problemas personales
¿Sientes que tienes tiempo suficiente para estu…
2

¿Cuántas horas dedicas diariamente a dormir?
2 a 6 hr
¿Cuántas horas dedicas diariamente a descansar?
3 a 5 hr
¿Estudias sol(a/o) o acompañad(a/o)?
Sol(a/o)
¿Si no sabes resolver un problema a quién acudes?
A un familiar
Por favor describe brevemente por qué estás est… Porque es una carrera que llevo hace años pens…
¿Has tenido la idea de retirarte o cambiarte a …
Sí
Solo si la respuesta anterior fue 'Sí' indica c… Cuando llega el limite de no tener tiempo por …

4
Marca temporal
9/12/2024 10:16:43
Dirección de correo electrónico
tamaratello6689@gmail.com
Nombres
Savska
Apellidos
Svec jimenez
Fecha de nacimiento
15/03/2002
¿Qué tanto te gusta estudiar?
3
¿Cuántas horas te dedicarías en una semana a es…
No estudiaría
¿Cuánto tiempo dedicas en una semana a estudiar… Estudio hasta que siento que estoy preparad(a/…
Si sientes que estás preparad(a/o) para una eva…
Definitivamente sí
Cuando estudias algo relacionado con Matemática…
Entre 10 y 20
¿Cuál o cuáles de los siguientes métodos utiliz…
Leer y subrayar
¿En qué horario prefieres estudiar? Me es indiferente (cualquier horario es bueno …
¿Qué lugar(es) utilizas para estudiar?
Una habitación
¿Qué factores consideras que dificultan tus est…
Estrés
¿Sientes que tienes tiempo suficiente para estu…
2
¿Cuántas horas dedicas diariamente a dormir?

```
6-8
¿Cuántas horas dedicas diariamente a descansar?
Todo el dia
¿Estudias sol(a/o) o acompañad(a/o)?
Sol(a/o)
¿Si no sabes resolver un problema a quién acudes?
NaN
Por favor describe brevemente por qué estás est…                    Por que
me gustan los negocios
¿Has tenido la idea de retirarte o cambiarte a …
Sí
Solo si la respuesta anterior fue 'Sí' indica c…
NaN
```

[4]:
```
encuesta.iloc[0,7]
```

[4]:
```
'Estudio hasta que siento que estoy preparad(a/o) no importa cuanto tiempo me
tome'
```

1.3 Valores NA

[5]:
```
encuesta.replace("", np.nan, inplace=True)
print(encuesta.isnull().sum())
encuesta.head(2).T
```

```
Marca temporal
0
Dirección de correo electrónico
0
Nombres
0
Apellidos
0
Fecha de nacimiento
0
¿Qué tanto te gusta estudiar?
0
¿Cuántas horas te dedicarías en una semana a estudiar para una asignatura si no
tuvieras una evaluación pronto?     0
¿Cuánto tiempo dedicas en una semana a estudiar para una asignatura en la que
pronto tendrás una evaluación?        0
Si sientes que estás preparad(a/o) para una evaluación ¿Dedicarías horas a
estudiar de todas formas?                0
Cuando estudias algo relacionado con Matemática…¿Cuántos ejercicios resuelves
en una sesión de estudio?           0
¿Cuál o cuáles de los siguientes métodos utilizas para estudiar?
0
¿En qué horario prefieres estudiar?
```

```
0
¿Qué lugar(es) utilizas para estudiar?
0
¿Qué factores consideras que dificultan tus estudios?
0
¿Sientes que tienes tiempo suficiente para estudiar?
0
¿Cuántas horas dedicas diariamente a dormir?
0
¿Cuántas horas dedicas diariamente a descansar?
0
¿Estudias sol(a/o) o acompañad(a/o)?
2
¿Si no sabes resolver un problema a quién acudes?
1
Por favor describe brevemente por qué estás estudiando esta carrera
3
¿Has tenido la idea de retirarte o cambiarte a otra carrera?
0
Solo si la respuesta anterior fue 'Sí' indica cuándo has pensado en cambiarte a
otra carrera o retirarte           43
dtype: int64
```

[5]:

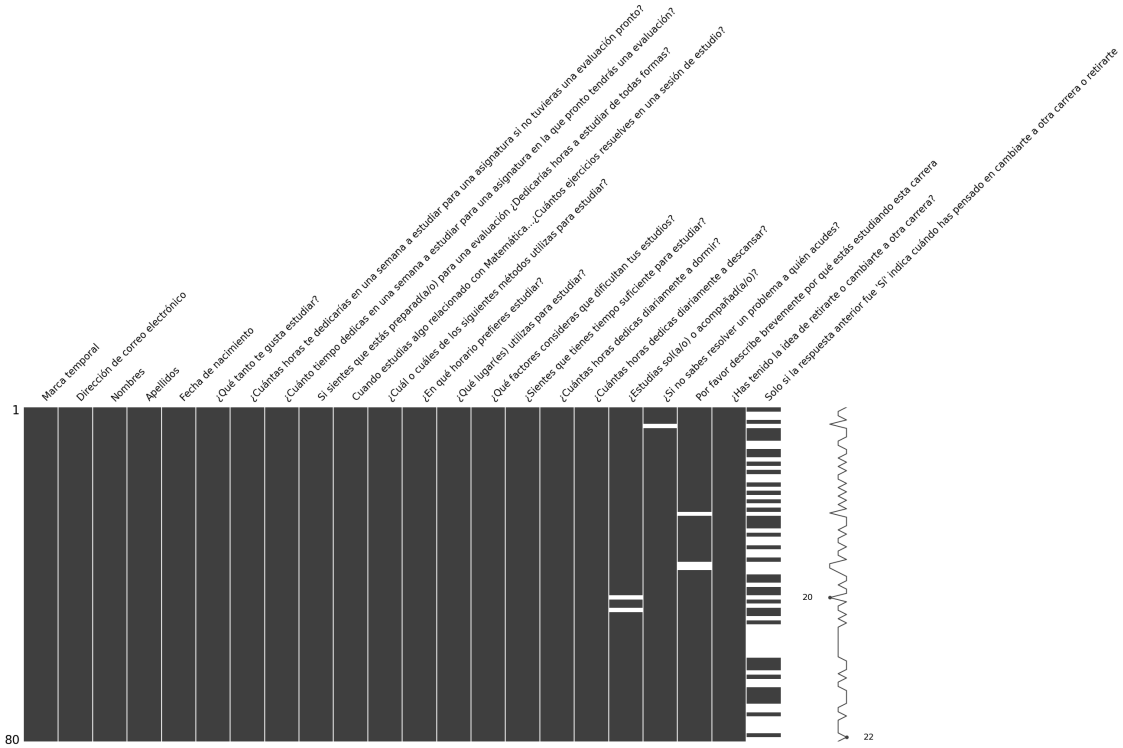
```
                     0  \
Marca temporal
9/12/2024 9:53:35
Dirección de correo electrónico
isidorahuaiquilaf402@gmail.com
Nombres
Isidora Esperanza
Apellidos
Huaiquilaf Verdugo
Fecha de nacimiento
3/10/2004
¿Qué tanto te gusta estudiar?
2
¿Cuántas horas te dedicarías en una semana a es…
Entre una y dos horas
¿Cuánto tiempo dedicas en una semana a estudiar…  Estudio hasta que siento que
estoy preparad(a/…
Si sientes que estás preparad(a/o) para una eva…
Definitivamente sí
Cuando estudias algo relacionado con Matemática…  Todos los necesarios (hasta
que siento que est…
¿Cuál o cuáles de los siguientes métodos utiliz…  Resolver ejercicios, Hacer
resúmenes, Elaborar…
¿En qué horario prefieres estudiar?                 En la mañana (desde que me
```

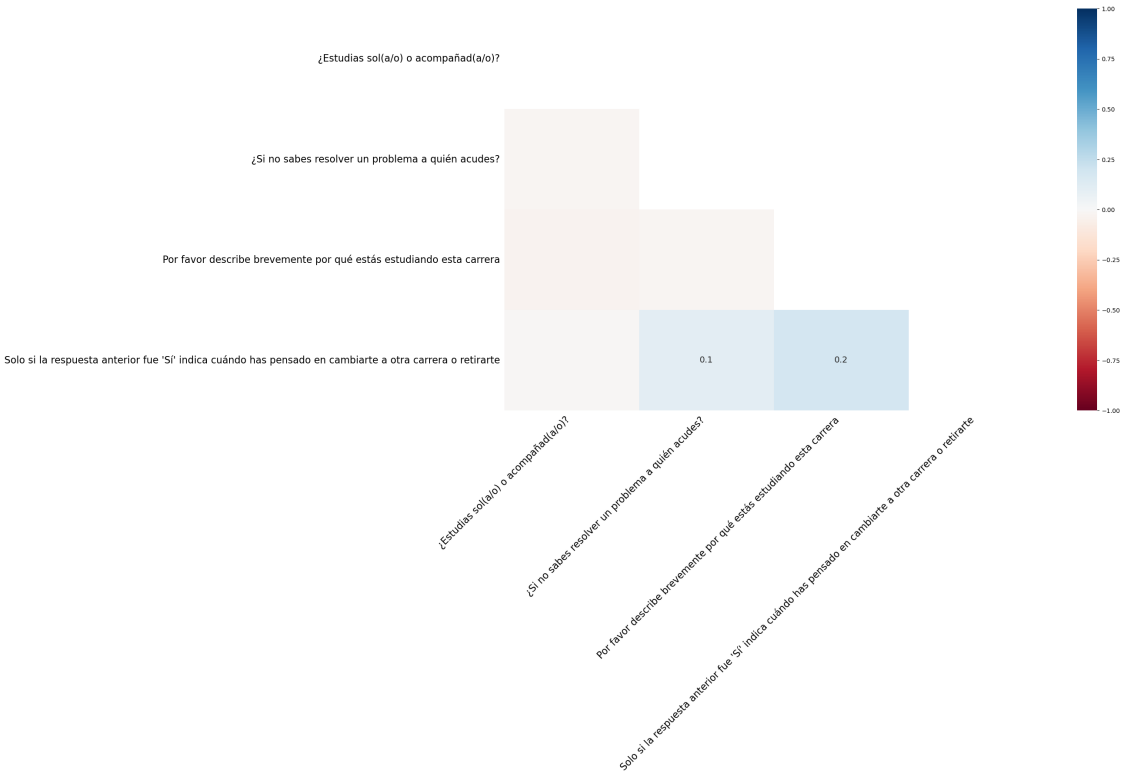
levanto hasta medio…
¿Qué lugar(es) utilizas para estudiar? Una habitación, Una pieza de estudio, Una bibl…
¿Qué factores consideras que dificultan tus est…
Estrés
¿Sientes que tienes tiempo suficiente para estu…
3
¿Cuántas horas dedicas diariamente a dormir?
6
¿Cuántas horas dedicas diariamente a descansar?
3
¿Estudias sol(a/o) o acompañad(a/o)?
Sol(a/o)
¿Si no sabes resolver un problema a quién acudes?
Profesor(a)
Por favor describe brevemente por qué estás est… Quiero saber constituir una empresa y me gusta…
¿Has tenido la idea de retirarte o cambiarte a …
Sí
Solo si la respuesta anterior fue 'Sí' indica c… Momentos de muchas pruebas y evaluaciones junt…

1
Marca temporal
9/12/2024 10:10:46
Dirección de correo electrónico
valentina.adm.ite@gmail.com
Nombres
Valentina del carmen
Apellidos
Silva ite
Fecha de nacimiento
13/07/2005
¿Qué tanto te gusta estudiar?
3
¿Cuántas horas te dedicarías en una semana a es…
Entre una y dos horas
¿Cuánto tiempo dedicas en una semana a estudiar…
Entre una y dos horas
Si sientes que estás preparad(a/o) para una eva…
Tal vez
Cuando estudias algo relacionado con Matemática… Ninguno.
Solo reviso el material
¿Cuál o cuáles de los siguientes métodos utiliz… Resolver ejercicios, Explicar los temas en voz…
¿En qué horario prefieres estudiar? En la noche (desde el atardecer hasta que me v…

```
¿Qué lugar(es) utilizas para estudiar?                                    Una
habitación, Una biblioteca
¿Qué factores consideras que dificultan tus est…
Falta de tiempo
¿Sientes que tienes tiempo suficiente para estu…
1
¿Cuántas horas dedicas diariamente a dormir?
5 horas
¿Cuántas horas dedicas diariamente a descansar?
Solo cuando duermo
¿Estudias sol(a/o) o acompañad(a/o)?
Sol(a/o)
¿Si no sabes resolver un problema a quién acudes?                          A un
amigo o amiga cercana
Por favor describe brevemente por qué estás est…  Por qué me gusta el
marketing y la relaciones …
¿Has tenido la idea de retirarte o cambiarte a …
No
Solo si la respuesta anterior fue 'Sí' indica c…
NaN
```

[6]:
```
# Visualizar los datos faltantes
msno.matrix(encuesta)
plt.show()
# Visualizar correlaciones entre datos faltantes
msno.heatmap(encuesta)
plt.show()
```


Marca temporal
Dirección de correo electrónico
Nombres
Apellidos
Fecha de nacimiento
¿Qué tanto te gusta estudiar?
¿Cuántas horas te dedicarías en una semana a estudiar para una asignatura si no tuvieras una evaluación pronto?
¿Cuánto tiempo dedicas en una semana a estudiar para una asignatura en la que pronto tendrás una evaluación?
Si sientes que estás preparad(a/o) para una evaluación ¿Dedicarías horas a estudiar de todas formas?
Cuando estudias algo relacionado con Matemática ¿Cuántos ejercicios resuelves en una sesión de estudio?
¿Cuál o cuáles de los siguientes métodos utilizas para estudiar?
¿En qué horario prefieres estudiar?
¿Qué lugar(es) utilizas para estudiar?
¿Qué factores consideras que dificultan tus estudios?
¿Sientes que tienes tiempo suficiente para estudiar?
¿Cuántas horas dedicas diariamente a dormir?
¿Cuántas horas dedicas diariamente a descansar?
¿Estudias sol(a/o) o acompañad(a/o)?
¿Si no sabes resolver un problema a quién acudes?
Por favor describe brevemente por qué estás estudiando esta carrera
¿Has tenido la idea de retirarte o cambiarte a otra carrera?
Solo si la respuesta anterior fue 'Si' indica cuándo has pensado en cambiarte a otra carrera o retirarte
1
80
20
22


¿Estudias sol(a/o) o acompañad(a/o)?
¿Si no sabes resolver un problema a quién acudes?
Por favor describe brevemente por qué estás estudiando esta carrera
Solo si la respuesta anterior fue 'Si' indica cuándo has pensado en cambiarte a otra carrera o retirarte
0.1
0.2
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

[7]:
```
# Example usage with your DataFrame `encuesta`
utils.test_missing_mcar(encuesta)
```

```
Chi2 Statistic: 0.0
P-value: 1.0
Degrees of Freedom: 0
Expected Frequencies: [36. 38.  3.  1.  1.  1.]
Result: Cannot reject the null hypothesis. The missing data is MCAR.
```

1.4 Homologación de valores

1.4.1 Transformar los valores de las preguntas

[8]:
```
# Load encodings
encodings = utils.load_yaml_encodings('../data/raw/
 ↪diccionario_codigos_respuestas.yaml')

# Process data
encoders = {}
processed_df = utils.process_survey_data(
    df=encuesta,
    encoding_dict=encodings,
    one_hot_encoders=encoders,
    ohe_categorical=False
)

#

processed_df.iloc[25,8] = 2
processed_df.iloc[8,9] = 3
processed_df.iloc[45,9] = 3

# Display the processed DataFrame
processed_df.head(2).T
```

[8]:
```
                    0  \
Marca temporal
9/12/2024 9:53:35
Dirección de correo electrónico
isidorahuaiquilaf402@gmail.com
Nombres
Isidora Esperanza
Apellidos
Huaiquilaf Verdugo
Fecha de nacimiento
3/10/2004
¿Qué tanto te gusta estudiar?
2
```

¿Cuántas horas te dedicarías en una semana a es…
3
¿Cuánto tiempo dedicas en una semana a estudiar…
6
Si sientes que estás preparad(a/o) para una eva…
5
Cuando estudias algo relacionado con Matemática…
5
¿Cuál o cuáles de los siguientes métodos utiliz…
A, B, C, D, F
¿En qué horario prefieres estudiar?
A, C
¿Qué lugar(es) utilizas para estudiar?
A, C, D
¿Qué factores consideras que dificultan tus est…
Estrés
¿Sientes que tienes tiempo suficiente para estu…
3
¿Cuántas horas dedicas diariamente a dormir?
6
¿Cuántas horas dedicas diariamente a descansar?
3
¿Estudias sol(a/o) o acompañad(a/o)?
A
¿Si no sabes resolver un problema a quién acudes?
A
Por favor describe brevemente por qué estás est… Quiero saber constituir una empresa y me gusta…
¿Has tenido la idea de retirarte o cambiarte a …
Sí
Solo si la respuesta anterior fue 'Sí' indica c… Momentos de muchas pruebas y evaluaciones junt…

	1
Marca temporal	9/12/2024 10:10:46
Dirección de correo electrónico	valentina.adm.ite@gmail.com
Nombres	Valentina del carmen
Apellidos	Silva ite
Fecha de nacimiento	13/07/2005
¿Qué tanto te gusta estudiar?	3
¿Cuántas horas te dedicarías en una semana a es…	

```
3
¿Cuánto tiempo dedicas en una semana a estudiar…
3
Si sientes que estás preparad(a/o) para una eva…
3
Cuando estudias algo relacionado con Matemática…
1
¿Cuál o cuáles de los siguientes métodos utiliz…
A, E
¿En qué horario prefieres estudiar?
C
¿Qué lugar(es) utilizas para estudiar?
A, D
¿Qué factores consideras que dificultan tus est…
Falta de tiempo
¿Sientes que tienes tiempo suficiente para estu…
1
¿Cuántas horas dedicas diariamente a dormir?
5 horas
¿Cuántas horas dedicas diariamente a descansar?
Solo cuando duermo
¿Estudias sol(a/o) o acompañad(a/o)?
A
¿Si no sabes resolver un problema a quién acudes?
B
Por favor describe brevemente por qué estás est…  Por qué me gusta el
marketing y la relaciones …
¿Has tenido la idea de retirarte o cambiarte a …
No
Solo si la respuesta anterior fue 'Sí' indica c…
NaN
```

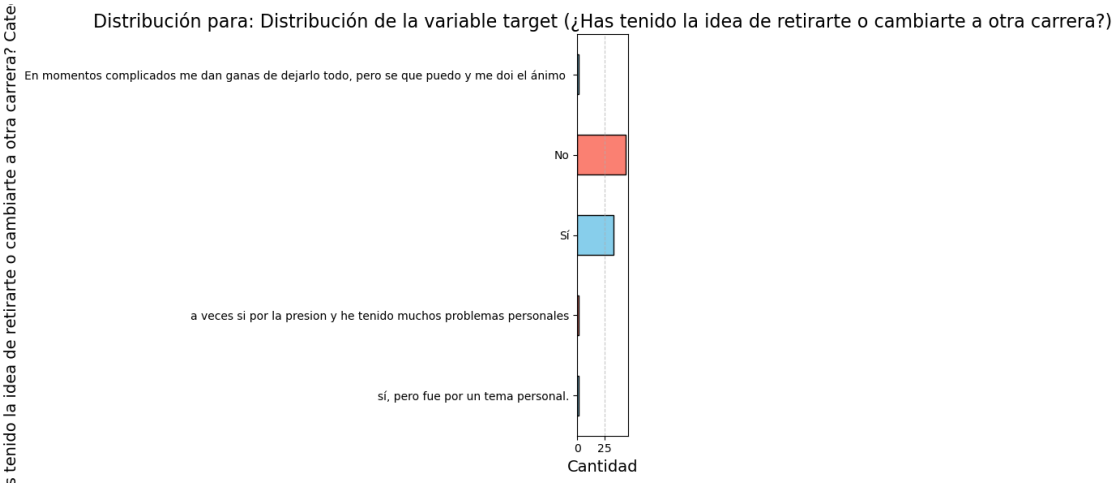
[9]:
```
encuesta.isnull().sum() == processed_df.isnull().sum()
```

[9]:
```
Marca temporal
True
Dirección de correo electrónico
True
Nombres
True
Apellidos
True
Fecha de nacimiento
True
¿Qué tanto te gusta estudiar?
True
¿Cuántas horas te dedicarías en una semana a estudiar para una asignatura si no
```

```
tuvieras una evaluación pronto?    True
¿Cuánto tiempo dedicas en una semana a estudiar para una asignatura en la que
pronto tendrás una evaluación?       True
Si sientes que estás preparad(a/o) para una evaluación ¿Dedicarías horas a
estudiar de todas formas?               True
Cuando estudias algo relacionado con Matemática…¿Cuántos ejercicios resuelves
en una sesión de estudio?          True
¿Cuál o cuáles de los siguientes métodos utilizas para estudiar?
True
¿En qué horario prefieres estudiar?
True
¿Qué lugar(es) utilizas para estudiar?
True
¿Qué factores consideras que dificultan tus estudios?
True
¿Sientes que tienes tiempo suficiente para estudiar?
True
¿Cuántas horas dedicas diariamente a dormir?
True
¿Cuántas horas dedicas diariamente a descansar?
True
¿Estudias sol(a/o) o acompañad(a/o)?
True
¿Si no sabes resolver un problema a quién acudes?
True
Por favor describe brevemente por qué estás estudiando esta carrera
True
¿Has tenido la idea de retirarte o cambiarte a otra carrera?
True
Solo si la respuesta anterior fue 'Sí' indica cuándo has pensado en cambiarte a
otra carrera o retirarte           True
dtype: bool
```

1.4.2 Homologación de valores en la variable target

[10]:
```python
# Call the function with P5 as an example
utils.plot_distribution(
    df=processed_df,
    variable_name="¿Has tenido la idea de retirarte o cambiarte a otra carrera?
 ↪",
    question_text='Distribución de la variable target',
    horizontal=True
)
```



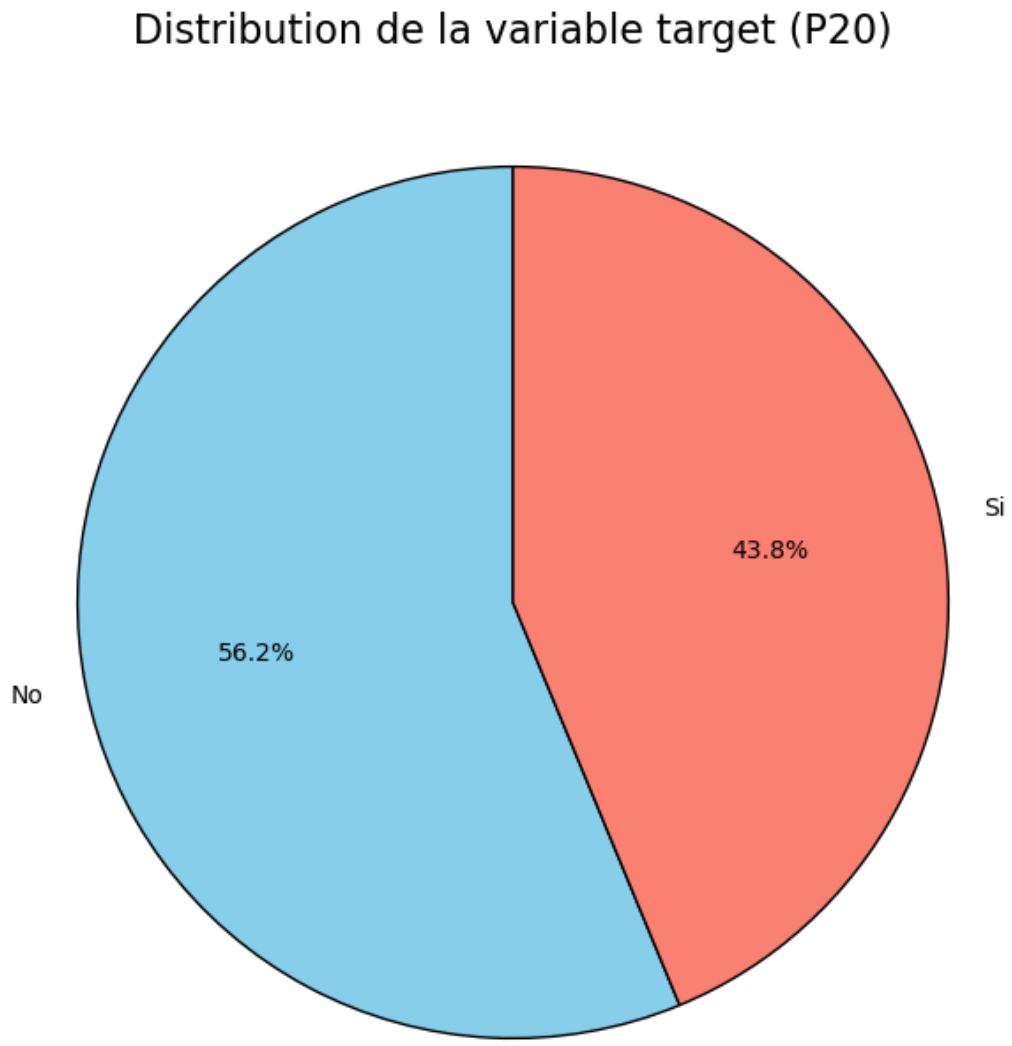
[11]:
```
# Normalize and transform P20
def transform_p20(response):
    # Normalize the text (lowercase, remove accents)
    normalized_response = unidecode.unidecode(str(response).strip().lower())

    # Return directly if already "sí" or "no"
    if normalized_response in ['si', 'no']:
        return normalized_response.capitalize()

    # Check for the presence of "si"
    if 'si' in normalized_response:
        return 'Si'
    else:
        return 'No'

# Apply transformation to P20
encuesta["¿Has tenido la idea de retirarte o cambiarte a otra carrera?"] =␣
 ↪encuesta["¿Has tenido la idea de retirarte o cambiarte a otra carrera?"].
 ↪apply(transform_p20)
# Apply transformation to processed_df
processed_df["¿Has tenido la idea de retirarte o cambiarte a otra carrera?"] =␣
 ↪encuesta["¿Has tenido la idea de retirarte o cambiarte a otra carrera?"].
 ↪map({'Si': 1, 'No': 0})
# Calculate the value counts and percentages
target_counts = encuesta["¿Has tenido la idea de retirarte o cambiarte a otra␣
 ↪carrera?"].value_counts(normalize=True) * 100
# Plot the pie chart
plt.figure(figsize=(8, 8))
plt.pie(
    target_counts,
```

```
    labels=target_counts.index,
    autopct='%1.1f%%',
    startangle=90,
    colors=['skyblue', 'salmon'],
    wedgeprops={'edgecolor': 'black'}
)
plt.title('Distribution de la variable target (P20)', fontsize=16)
plt.show()
```



1.4.3 Calcular la fecha de nacimiento

[12]:
```
processed_df['Fecha de nacimiento'] = processed_df['Fecha de nacimiento'].
↪apply(utils.calculate_age)
```

1.5 Homologar nombres de variables

[13]:
```
# Load the JSON file
with open(__path__ + 'diccionario_nombre_variables.json', 'r') as file:
    diccionario = json.load(file)

# Extract the mapping (preguntas -> P codes)
preguntas_mapping = {v['pregunta']: k for k, v in diccionario.items()}
# Rename columns
encuesta.rename(columns=preguntas_mapping, inplace=True)
print(encuesta.info())
encuesta.head(3).T
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 80 entries, 0 to 79
Data columns (total 22 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   P0      80 non-null     object
 1   P1      80 non-null     object
 2   P2      80 non-null     object
 3   P3      80 non-null     object
 4   P4      80 non-null     object
 5   P5      80 non-null     int64
 6   P6      80 non-null     object
 7   P7      80 non-null     object
 8   P8      80 non-null     object
 9   P9      80 non-null     object
 10  P10     80 non-null     object
 11  P11     80 non-null     object
 12  P12     80 non-null     object
 13  P13     80 non-null     object
 14  P14     80 non-null     int64
 15  P15     80 non-null     object
 16  P16     80 non-null     object
 17  P17     78 non-null     object
 18  P18     79 non-null     object
 19  P19     77 non-null     object
 20  P20     80 non-null     object
 21  P21     37 non-null     object
dtypes: int64(2), object(20)
memory usage: 13.9+ KB
None
```

[13]:

	0	\
P0	9/12/2024 9:53:35	
P1	isidorahuaiquilaf402@gmail.com	
P2	Isidora Esperanza	
P3	Huaiquilaf Verdugo	
P4	3/10/2004	
P5	2	
P6	Entre una y dos horas	
P7	Estudio hasta que siento que estoy preparad(a/…	
P8	Definitivamente sí	
P9	Todos los necesarios (hasta que siento que est…	
P10	Resolver ejercicios, Hacer resúmenes, Elaborar…	
P11	En la mañana (desde que me levanto hasta medio…	
P12	Una habitación, Una pieza de estudio, Una bibl…	
P13	Estrés	
P14	3	
P15	6	
P16	3	
P17	Sol(a/o)	
P18	Profesor(a)	
P19	Quiero saber constituir una empresa y me gusta…	
P20	Si	
P21	Momentos de muchas pruebas y evaluaciones junt…	

	1	\
P0	9/12/2024 10:10:46	
P1	valentina.adm.ite@gmail.com	
P2	Valentina del carmen	
P3	Silva ite	
P4	13/07/2005	
P5	3	
P6	Entre una y dos horas	
P7	Entre una y dos horas	
P8	Tal vez	
P9	Ninguno. Solo reviso el material	
P10	Resolver ejercicios, Explicar los temas en voz…	
P11	En la noche (desde el atardecer hasta que me v…	
P12	Una habitación, Una biblioteca	
P13	Falta de tiempo	
P14	1	
P15	5 horas	
P16	Solo cuando duermo	
P17	Sol(a/o)	
P18	A un amigo o amiga cercana	
P19	Por qué me gusta el marketing y la relaciones …	
P20	No	
P21	NaN	

```
                                                    2
P0                                   9/12/2024 10:12:27
P1                         aljaque@alumnos.uahurtado.cl
P2                                      Alexis Alexander
P3                                        Jaque Cardenas
P4                                            30/06/2006
P5                                                    4
P6   Estudiaría hasta que sienta que estoy preparad…
P7                                     Entre 4 y 7 horas
P8                                    Definitivamente sí
P9   Todos los necesarios (hasta que siento que est…
P10  Resolver ejercicios, Hacer resúmenes, Elaborar…
P11  Me es indiferente (cualquier horario es bueno …
P12  Una habitación, Una pieza de estudio, Una bibl…
P13  Puede ser falta de recursos y tiempo, ya que t…
P14                                                   3
P15                                            10 u 8
P16                                           2 horas
P17                                          Sol(a/o)
P18                                        Profesor(a)
P19          Porque me llama la atención el comercio
P20                                                  No
P21                                                 NaN
```

[14]:
```python
# Load the JSON file
with open(__path__ + 'diccionario_nombre_variables.json', 'r') as file:
    diccionario = json.load(file)

# Extract the mapping (preguntas -> P codes)
preguntas_mapping = {v['pregunta']: k for k, v in diccionario.items()}
# Rename columns
processed_df.rename(columns=preguntas_mapping, inplace=True)
print(processed_df.info())
processed_df.head(3).T
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 80 entries, 0 to 79
Data columns (total 22 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   P0      80 non-null     object
 1   P1      80 non-null     object
 2   P2      80 non-null     object
 3   P3      80 non-null     object
 4   P4      80 non-null     int64
 5   P5      80 non-null     int64
 6   P6      80 non-null     int64
```

```
 7   P7      80 non-null     int64
 8   P8      80 non-null     object
 9   P9      80 non-null     object
 10  P10     80 non-null     object
 11  P11     80 non-null     object
 12  P12     80 non-null     object
 13  P13     80 non-null     object
 14  P14     80 non-null     int64
 15  P15     80 non-null     object
 16  P16     80 non-null     object
 17  P17     78 non-null     object
 18  P18     79 non-null     object
 19  P19     77 non-null     object
 20  P20     80 non-null     int64
 21  P21     37 non-null     object
dtypes: int64(6), object(16)
memory usage: 13.9+ KB
None
```

[14]:

	0	1
P0	9/12/2024 9:53:35	9/12/2024 10:10:46
P1	isidorahuaiquilaf402@gmail.com	valentina.adm.ite@gmail.com
P2	Isidora Esperanza	Valentina del carmen
P3	Huaiquilaf Verdugo	
P4	20	
P5	2	
P6	3	
P7	6	
P8	5	
P9	5	
P10	A, B, C, D, F	
P11	A, C	
P12	A, C, D	
P13	Estrés	
P14	3	
P15	6	
P16	3	
P17	A	
P18	A	
P19	Quiero saber constituir una empresa y me gusta…	
P20	1	
P21	Momentos de muchas pruebas y evaluaciones junt…	

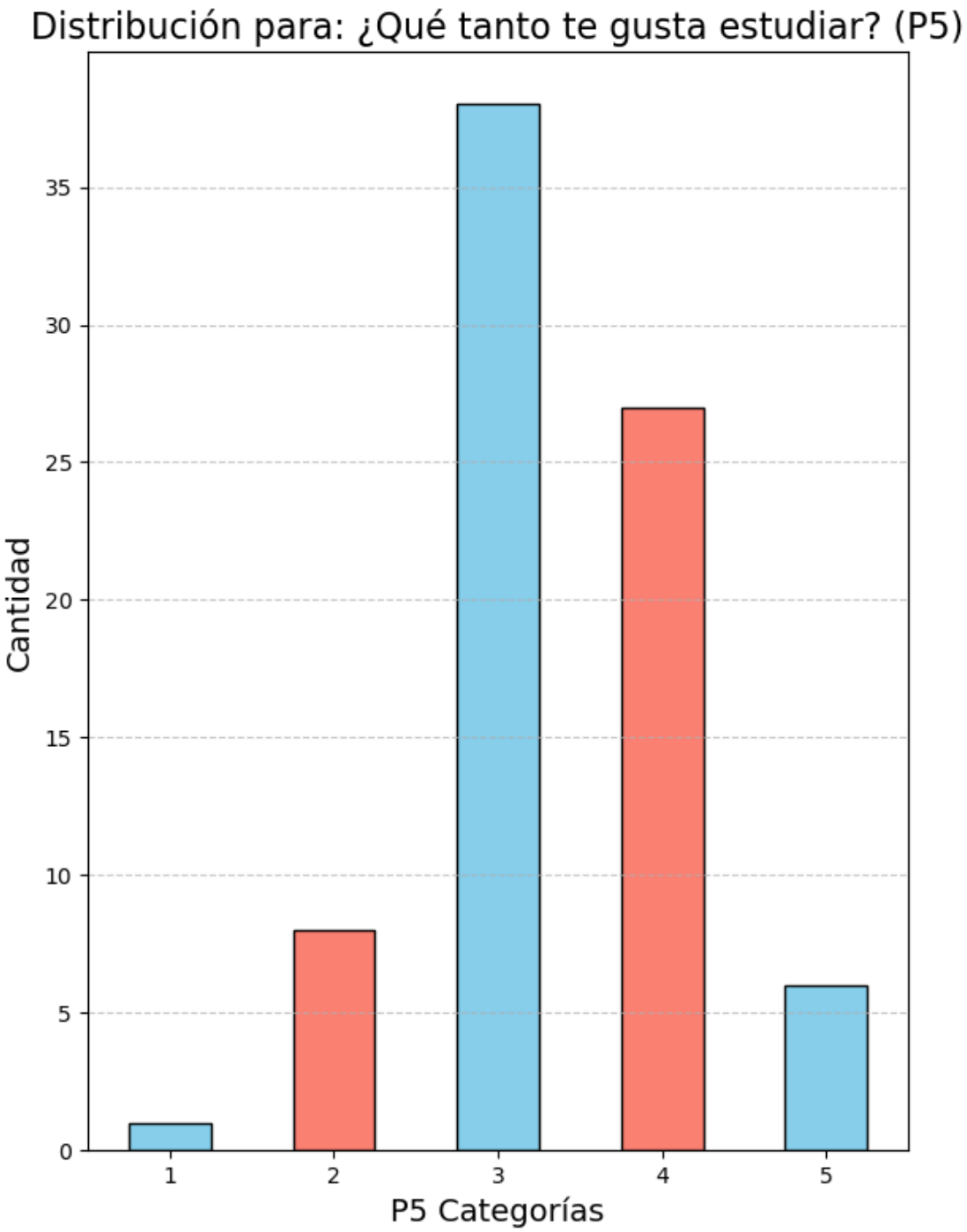
```
P3                                              Silva ite
P4                                                     19
P5                                                      3
P6                                                      3
P7                                                      3
P8                                                      3
P9                                                      1
P10                                                A, E
P11                                                    C
P12                                                A, D
P13                                    Falta de tiempo
P14                                                     1
P15                                             5 horas
P16                                  Solo cuando duermo
P17                                                    A
P18                                                    B
P19  Por qué me gusta el marketing y la relaciones …
P20                                                     0
P21                                                   NaN

                                                        2
P0                                   9/12/2024 10:12:27
P1                          aljaque@alumnos.uahurtado.cl
P2                                     Alexis Alexander
P3                                       Jaque Cardenas
P4                                                     18
P5                                                      4
P6                                                      6
P7                                                      5
P8                                                      5
P9                                                      5
P10                                         A, B, C, F
P11                                                    D
P12                                            A, C, D
P13  Puede ser falta de recursos y tiempo, ya que t…
P14                                                     3
P15                                             10 u 8
P16                                             2 horas
P17                                                    A
P18                                                    A
P19           Porque me llama la atención el comercio
P20                                                     0
P21                                                   NaN
```

1.6 Análisis bi-variable con respecto al target

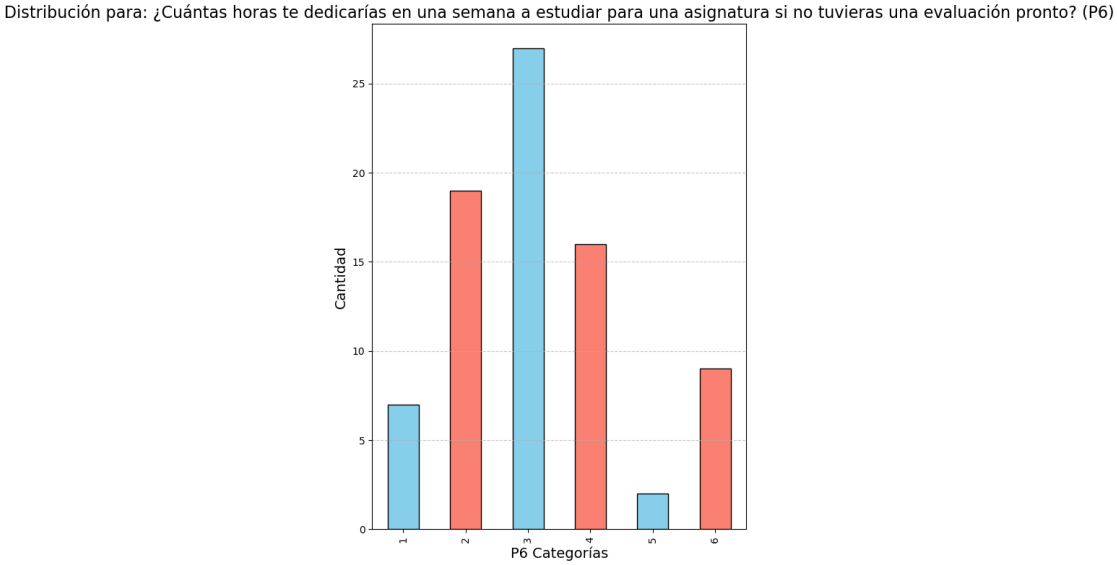
Distribución para las preguntas de hábitos de estudio

[15]:
```
# Call the function with P5 as an example
utils.plot_distribution(
    df=processed_df,
    variable_name='P5',
    question_text='¿Qué tanto te gusta estudiar?',
    horizontal=False,
    ascending=True
)
```



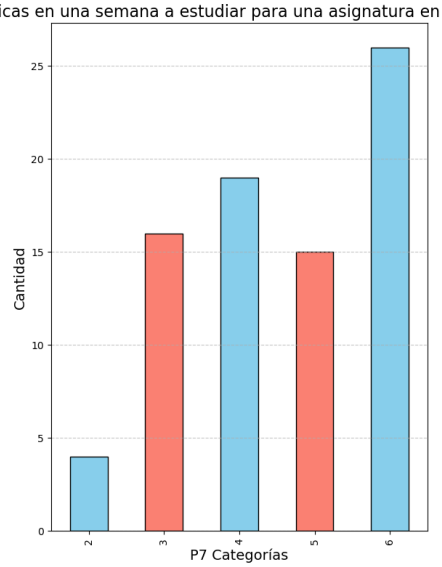
[16]:
```
# Call the function with P5 as an example
utils.plot_distribution(
    df=processed_df,
    variable_name='P6',
```

```
    question_text='¿Cuántas horas te dedicarías en una semana a estudiar para una asignatura si no tuvieras una evaluación pronto?',
    rotation=90,
    horizontal=False,
    ascending=True
    #figsize=(20,6)
)
```



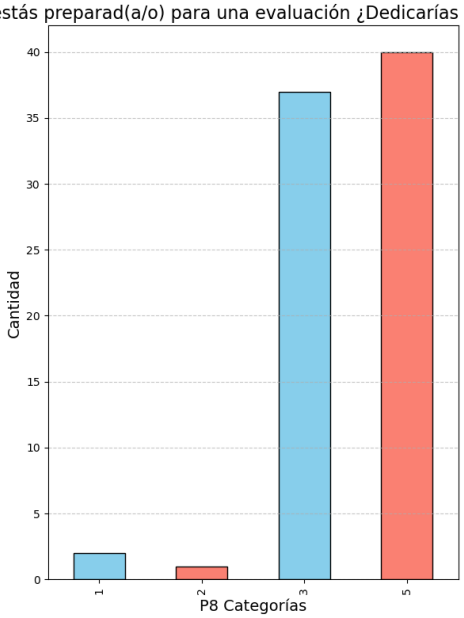
[17]:
```
# Call the function with P5 as an example
utils.plot_distribution(
    df=processed_df,
    variable_name='P7',
    question_text="¿Cuánto tiempo dedicas en una semana a estudiar para una asignatura en la que pronto tendrás una evaluación?",
    rotation=90,
    horizontal=False,
    ascending=True
    #figsize=(20,6)
)
```

Distribución para: ¿Cuánto tiempo dedicas en una semana a estudiar para una asignatura en la que pronto tendrás una evaluación? (P7)



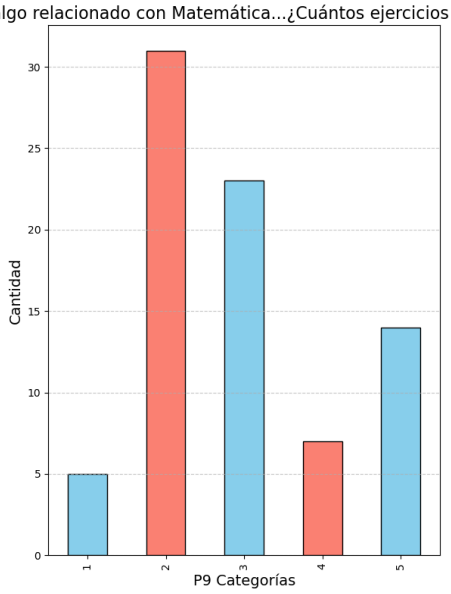
[18]:
```
# Call the function with P5 as an example
utils.plot_distribution(
    df=processed_df,
    variable_name='P8',
    question_text="Si sientes que estás preparad(a/o) para una evaluación␣
 ↪¿Dedicarías horas a estudiar de todas formas?",
    rotation=90,
    horizontal=False,
    ascending=True
    #figsize=(20,6)
)
```

Distribución para: Si sientes que estás preparad(a/o) para una evaluación ¿Dedicarías horas a estudiar de todas formas? (P8)



[19]:
```
utils.plot_distribution(
    df=processed_df,
    variable_name='P9',
    question_text= "Cuando estudias algo relacionado con Matemática...¿Cuántos␣
 ↪ejercicios resuelves en una sesión de estudio?",
    rotation=90,
    horizontal=False,
    ascending=True
    #figsize=(20,6)
)
```

Distribución para: Cuando estudias algo relacionado con Matemática...¿Cuántos ejercicios resuelves en una sesión de estudio? (P9)



1.6.1 Test estadísticos con variables asociadas a hábitos de estudio

[34]:
```
from scipy import stats

def rank_biserial_correlation(likert_var, binary_var):
    return stats.pointbiserialr(binary_var, likert_var)[0]

def interpret_rank_biserial(correlation):
   """
   Interprets rank-biserial correlation coefficient.
   """
   strength = ""
   if abs(correlation) < 0.2: strength = "Very weak"
   elif abs(correlation) < 0.4: strength = "Weak"
   elif abs(correlation) < 0.6: strength = "Moderate"
   elif abs(correlation) < 0.8: strength = "Strong"
   else: strength = "Very strong"

   direction = "positive" if correlation > 0 else "negative"

   return f"Correlation: {correlation:.3f}\nStrength: {strength}\nDirection:␣
 ↪{direction}"

# Convert columns P4-P9 and P14 to int64
columns_to_convert = [f'P{i}' for i in range(4, 10)] + ['P14']
processed_df[columns_to_convert] = processed_df[columns_to_convert].
 ↪astype('int64')
```

```
for question in range(5, 10):
    col_name = f'P{question}'
    if col_name in processed_df.columns:
        result = rank_biserial_correlation(processed_df[col_name].values,␣
 ↪processed_df.P20.values)
        print(f"\nQuestion {col_name}:")
        print(interpret_rank_biserial(result))
result = rank_biserial_correlation(processed_df['P14'].values, processed_df.P20.
 ↪values)
print(f"\nQuestion {'P14'}:")
print(interpret_rank_biserial(result))
```

```
Question P5:
Correlation: -0.270
Strength: Weak
Direction: negative

Question P6:
Correlation: -0.316
Strength: Weak
Direction: negative

Question P7:
Correlation: -0.116
Strength: Very weak
Direction: negative

Question P8:
Correlation: -0.018
Strength: Very weak
Direction: negative

Question P9:
Correlation: 0.204
Strength: Weak
Direction: positive

Question P14:
Correlation: 0.048
Strength: Very weak
Direction: positive
```

[35]:
```
from scipy import stats

def mann_whitney_test(likert_var, binary_var):
    group_0 = likert_var[binary_var == 0]
```

```
    group_1 = likert_var[binary_var == 1]
    return stats.mannwhitneyu(group_0, group_1, alternative='two-sided')

def interpret_mann_whitney(statistic, pvalue, n1, n2):
    """
    Interprets Mann-Whitney U test results
    """
    effect_size = 1 - (2 * statistic)/(n1 * n2)  # Common language effect size

    strength = ""
    if abs(effect_size) < 0.2: strength = "Very weak"
    elif abs(effect_size) < 0.4: strength = "Weak"
    elif abs(effect_size) < 0.6: strength = "Moderate"
    elif abs(effect_size) < 0.8: strength = "Strong"
    else: strength = "Very strong"

    significance = "significant" if pvalue < 0.05 else "not significant"

    return f"Effect size: {effect_size:.3f}\nStrength: {strength}\np-value:␣
 ↪{pvalue:.3f}\nStatistical significance: {significance}"

# Convert columns
columns_to_convert = [f'P{i}' for i in range(4, 10)] + ['P14']
processed_df[columns_to_convert] = processed_df[columns_to_convert].
 ↪astype('int64')

# Run tests
for question in range(5, 10):
    col_name = f'P{question}'
    if col_name in processed_df.columns:
        stat, pval = mann_whitney_test(processed_df[col_name].values,␣
 ↪processed_df.P20.values)
        n1 = sum(processed_df.P20 == 0)
        n2 = sum(processed_df.P20 == 1)
        print(f"\nQuestion {col_name}:")
        print(interpret_mann_whitney(stat, pval, n1, n2))

# Test P14
stat, pval = mann_whitney_test(processed_df['P14'].values, processed_df.P20.
 ↪values)
n1 = sum(processed_df.P20 == 0)
n2 = sum(processed_df.P20 == 1)
print(f"\nQuestion P14:")
print(interpret_mann_whitney(stat, pval, n1, n2))
```

```
Question P5:
```

```
Effect size: -0.277
Strength: Weak
p-value: 0.022
Statistical significance: significant

Question P6:
Effect size: -0.363
Strength: Weak
p-value: 0.004
Statistical significance: significant

Question P7:
Effect size: -0.108
Strength: Very weak
p-value: 0.397
Statistical significance: not significant

Question P8:
Effect size: -0.017
Strength: Very weak
p-value: 0.886
Statistical significance: not significant

Question P9:
Effect size: 0.221
Strength: Weak
p-value: 0.078
Statistical significance: not significant

Question P14:
Effect size: 0.035
Strength: Very weak
p-value: 0.780
Statistical significance: not significant
```