No-Show Medical Appointments Analysis

Brazil 2016 - 110,527 appointments

Goal: Identify factors that predict whether patients show up (No-show = Yes)

Submitted by: Jerad Williams - 2/24/2026

In [29]:
```
!pip install matplotlib seaborn
```

```
Requirement already satisfied: matplotlib in /Library/Frameworks/Python.fram
ework/Versions/3.14/lib/python3.14/site-packages (3.10.8)
Requirement already satisfied: seaborn in /Library/Frameworks/Python.framewo
rk/Versions/3.14/lib/python3.14/site-packages (0.13.2)
Requirement already satisfied: contourpy>=1.0.1 in /Library/Frameworks/Pytho
n.framework/Versions/3.14/lib/python3.14/site-packages (from matplotlib) (1.
3.3)
Requirement already satisfied: cycler>=0.10 in /Library/Frameworks/Python.fr
amework/Versions/3.14/lib/python3.14/site-packages (from matplotlib) (0.12.
1)
Requirement already satisfied: fonttools>=4.22.0 in /Library/Frameworks/Pyth
on.framework/Versions/3.14/lib/python3.14/site-packages (from matplotlib)
(4.61.1)
Requirement already satisfied: kiwisolver>=1.3.1 in /Library/Frameworks/Pyth
on.framework/Versions/3.14/lib/python3.14/site-packages (from matplotlib)
(1.4.9)
Requirement already satisfied: numpy>=1.23 in /Library/Frameworks/Python.fra
mework/Versions/3.14/lib/python3.14/site-packages (from matplotlib) (2.4.2)
Requirement already satisfied: packaging>=20.0 in /Library/Frameworks/Pytho
n.framework/Versions/3.14/lib/python3.14/site-packages (from matplotlib) (2
6.0)
Requirement already satisfied: pillow>=8 in /Library/Frameworks/Python.frame
work/Versions/3.14/lib/python3.14/site-packages (from matplotlib) (12.1.1)
Requirement already satisfied: pyparsing>=3 in /Library/Frameworks/Python.fr
amework/Versions/3.14/lib/python3.14/site-packages (from matplotlib) (3.3.2)
Requirement already satisfied: python-dateutil>=2.7 in /Library/Frameworks/P
ython.framework/Versions/3.14/lib/python3.14/site-packages (from matplotlib)
(2.9.0.post0)
Requirement already satisfied: pandas>=1.2 in /Library/Frameworks/Python.fra
mework/Versions/3.14/lib/python3.14/site-packages (from seaborn) (3.0.1)
Requirement already satisfied: six>=1.5 in /Library/Frameworks/Python.framew
ork/Versions/3.14/lib/python3.14/site-packages (from python-dateutil>=2.7->m
atplotlib) (1.17.0)
```

In [30]:
```
!ls
```

```
airflow
Applications
dbt_healthcare
Desktop
Documents
Downloads
healthcare-patient-analytics-pipeline
healthcare-venv
Library
linkedin photo.jpeg
Movies
Music
NoShow_Analysis_Final.ipynb
noshowappointments-kagglev2-may-2016.csv
Pictures
Public
```

In [31]:

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline

df = pd.read_csv('noshowappointments-kagglev2-may-2016.csv')
print('Shape:', df.shape)
df.head()
```

```
Shape: (110527, 14)
```

Out[31]:

	PatientId	AppointmentID	Gender	ScheduledDay	AppointmentDay	Age	Neig
0	2.987250e+13	5642903	F	2016-04-29T18:38:08Z	2016-04-29T00:00:00Z	62	
1	5.589978e+14	5642503	M	2016-04-29T16:08:27Z	2016-04-29T00:00:00Z	56	
2	4.262962e+12	5642549	F	2016-04-29T16:19:04Z	2016-04-29T00:00:00Z	62	MAT.
3	8.679512e+11	5642828	F	2016-04-29T17:29:31Z	2016-04-29T00:00:00Z	8	I
4	8.841186e+12	5642494	F	2016-04-29T16:07:23Z	2016-04-29T00:00:00Z	56	

In [32]:

```python
df = df.rename(columns={'Hipertension':'Hypertension', 'Handcap':'Handicap',
df['NoShow'] = (df['NoShow'] == 'Yes').astype(int)

df['ScheduledDay'] = pd.to_datetime(df['ScheduledDay'])
df['AppointmentDay'] = pd.to_datetime(df['AppointmentDay'])
df['WaitDays'] = (df['AppointmentDay'] - df['ScheduledDay']).dt.days
df['AgeGroup'] = pd.cut(df['Age'], bins=[-1,18,35,55,120], labels=['Child','
df['HasChronic'] = ((df['Hypertension'] + df['Diabetes']) > 0).astype(int)

print('Overall no-show rate: {:.1%}'.format(df['NoShow'].mean()))
```

```
Overall no-show rate: 20.2%
```

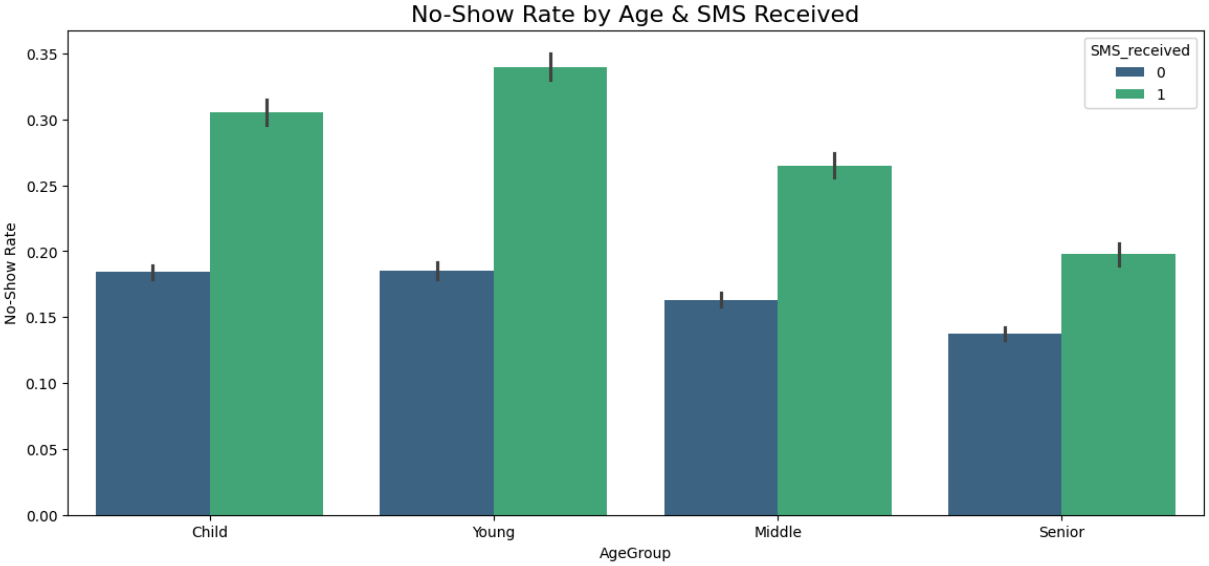
Research Questions

1. How do **SMS**, **AgeGroup**, and **WaitDays** interact?
2. Does **Scholarship** interact with **Age** and **Chronic conditions**?
3. Which **Neighbourhoods** have highest risk?

In [33]:

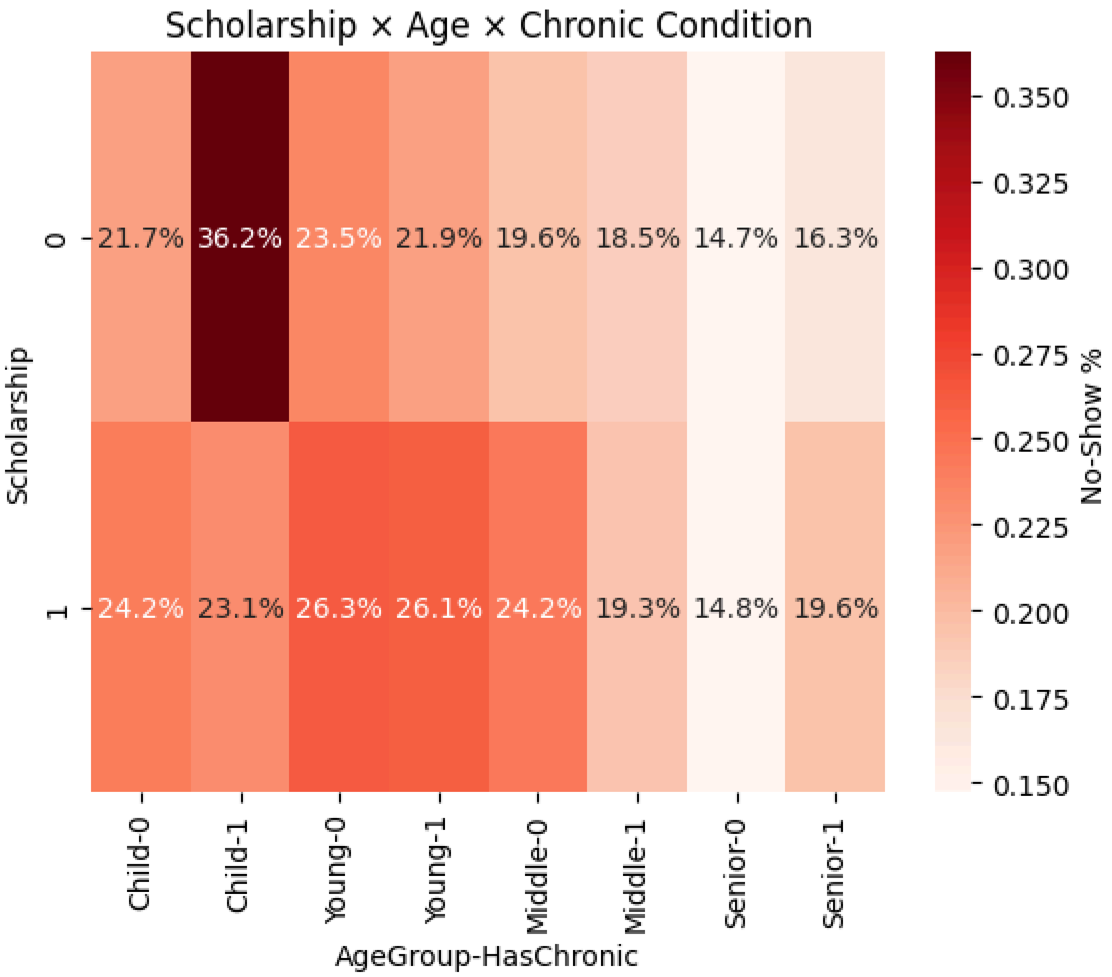
```
df['WaitGroup'] = pd.cut(df['WaitDays'], bins=[-1,0,7,30,180], labels=['Same

plt.figure(figsize=(14,6))
sns.barplot(data=df, x='AgeGroup', y='NoShow', hue='SMS_received', palette='
plt.title('No-Show Rate by Age & SMS Received', fontsize=16)
plt.ylabel('No-Show Rate')
plt.show()
```



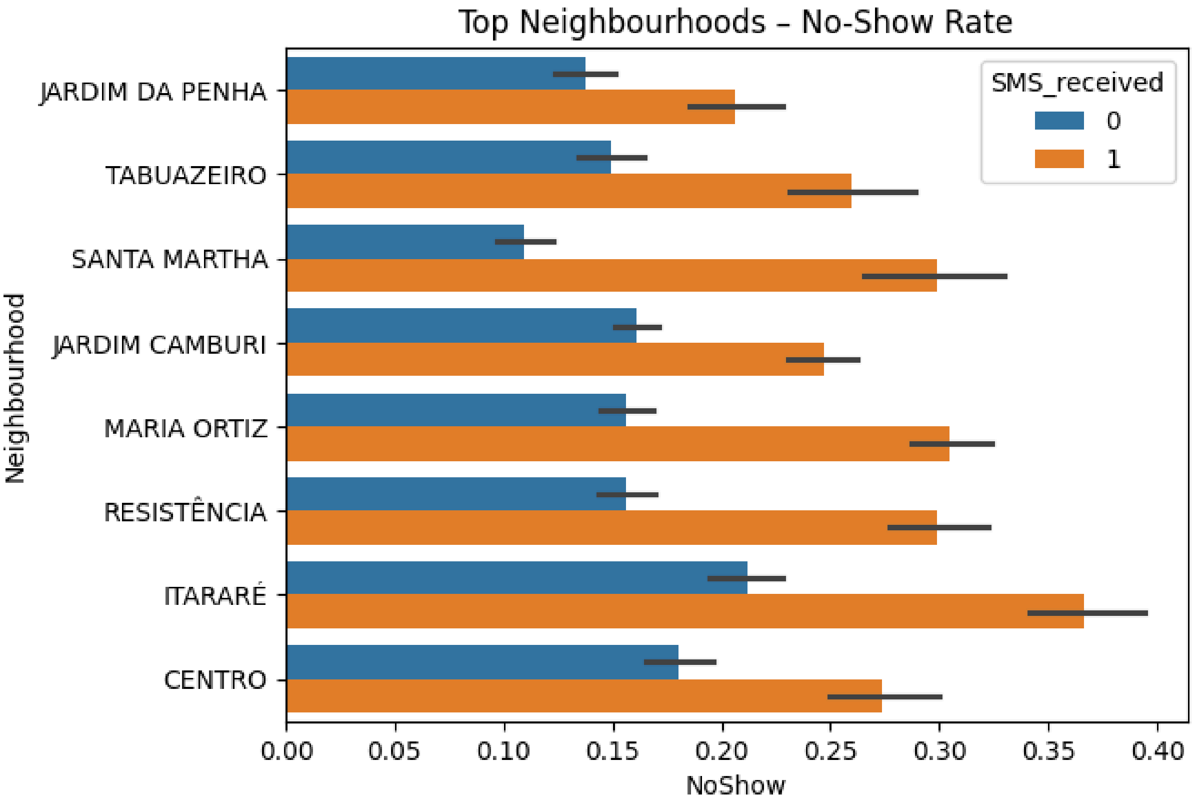
In [34]:

```
pivot = df.pivot_table('NoShow', index='Scholarship', columns=['AgeGroup','H
sns.heatmap(pivot, annot=True, fmt='.1%', cmap='Reds', cbar_kws={'label': 'N
plt.title('Scholarship × Age × Chronic Condition')
plt.show()
```



In [35]:

```
top_n = df['Neighbourhood'].value_counts().head(8).index
top_df = df[df['Neighbourhood'].isin(top_n)]
sns.barplot(data=top_df, y='Neighbourhood', x='NoShow', hue='SMS_received')
plt.title('Top Neighbourhoods – No-Show Rate')
plt.show()
```



Key Findings & Recommendations

- **Strongest factor:** Wait time >30 days → 30%+ no-show
- **Highest risk:** Young adults (15-35) with long waits + no SMS (34%)
- **SMS helps** long-wait patients but is sent to higher-risk groups
- Scholarship patients show up less, especially young & healthy
- Neighbourhoods vary widely → target high-risk areas first

Dependent variable: NoShow
Independent variables: SMS_received, Age, WaitDays, Scholarship, HasChronic, Neighbourhood