Exploring Factors Influencing Patient Attendance in Medical Appointments

INDEX

1. Introduction

1.1 Background

In healthcare systems, missed medical appointments, commonly referred to as no-shows, represent a significant challenge. These no-shows not only disrupt the scheduling efficiency but also lead to wasted resources and impact the quality of patient care. Understanding the factors contributing to such behavior is crucial for healthcare organizations to optimize their operations and ensure better patient outcomes.

The dataset for this study contains medical appointment records, providing insights into factors such as patient demographics, medical conditions, and appointment details. By leveraging statistical techniques and SAS for exploratory data analysis, this project aims to identify patterns and relationships that influence whether a patient attends their scheduled appointment. Such insights can aid in designing strategies to reduce no-shows and improve healthcare delivery.

1.2 Objectives

The primary objectives of this project are:

- Determine the percentage of appointments where patients failed to show up.
- Explore how various factors such as gender, age, medical conditions, and socioeconomic status influence patient attendance.
- Assess the effect of sending SMS reminders on patient attendance behavior.
- Study the relationship between the time gap (scheduling vs. appointment day) and attendance likelihood.
- Identify neighborhoods with the highest no-show rates.
- Determine the weekdays with the most missed appointments.
- Draw conclusions and suggest strategies to reduce no-show rates and enhance appointment management systems.

2. Data Description

2.1 Dataset Overview

The dataset used for this study comprises medical appointment records and includes a total of 110,527 observations with 14 variables. Each record represents an individual medical appointment, providing information about patient demographics, appointment details, and attendance status.

This data enables an in-depth analysis of patient behavior, highlighting patterns and factors associated with missed appointments (no-shows). The data was originally structured in a CSV format and was imported into the SAS environment for analysis.

2.2 Data Dictionary

a. PatientId: Unique identifier for each patient.
b. AppointmentID: Unique identifier for each appointment.
c. Gender: Patient's gender (Male or Female).
d. AppointmentDay: Date of the scheduled appointment.
e. ScheduledDay: Date when the appointment was booked.
f. Age: Patient's age (including negative values indicating unborn children).
g. Neighbourhood: Location where the appointment was scheduled.
h. Scholarship: Indicates whether the patient is on a government health assistance program (1 = Yes, 0 = No).
i. Hypertension: Indicates if the patient has hypertension (1 = Yes, 0 = No).
j. Diabetes: Indicates if the patient has diabetes (1 = Yes, 0 = No).
k. Alcoholism: Indicates if the patient has a history of alcohol dependence (1 = Yes, 0 = No).
l. Handicap: Indicates if the patient has a disability (1 = Yes, 0 = No).
m. SMS_received: Number of SMS reminders sent to the patient (1 = At least one, 0 = None).
n. No-show: Indicates if the patient missed the appointment (Yes = Missed, No = Attended).

2.3 Notes on Data Collection

The dataset was collected from healthcare facilities and represents real-world scenarios of patient attendance and scheduling behavior.

The data includes several observations for the same patients, reflecting repeated appointments.

The No-show column was inverted and renamed to Show for ease of interpretation, where 1 = Attended and 0 = Missed.

The dataset has no missing values, which simplifies data preprocessing. However, certain variables, such as Age, contain negative values representing unborn children.

The dataset does not specify the exact appointment times or whether certain days were holidays, which could influence patient attendance.

The data used in this project is taken from a guided project (Link: https://cognitiveclass.ai/courses/medical-appointment-data-analysis).

The original analysis is done using Python and its modules such as numpy, matplotlib, seaborn, and pandas.

Data can also be downloaded from Kaggle website (Link: https://www.kaggle.com/datasets/joniarroba/noshowappointments?resource=download&select=KaggleV2-May-2016.csv)

3. Methodology

3.1 Data Setup in SAS

The dataset was provided in .csv format, containing 110,527 rows and 14 columns. To begin, the file was uploaded to the SAS Studio folder named Medical Appointments. The PROC IMPORT procedure was utilized to load the dataset into the SAS environment. This step ensured that all columns and data types were accurately imported for further analysis.

Key steps included:

- Specifying the file path and defining the data structure.
- Checking the successful import of data by inspecting the first few rows using PROC PRINT.
- Confirming the column names, data types, and the number of rows and columns using PROC CONTENTS.

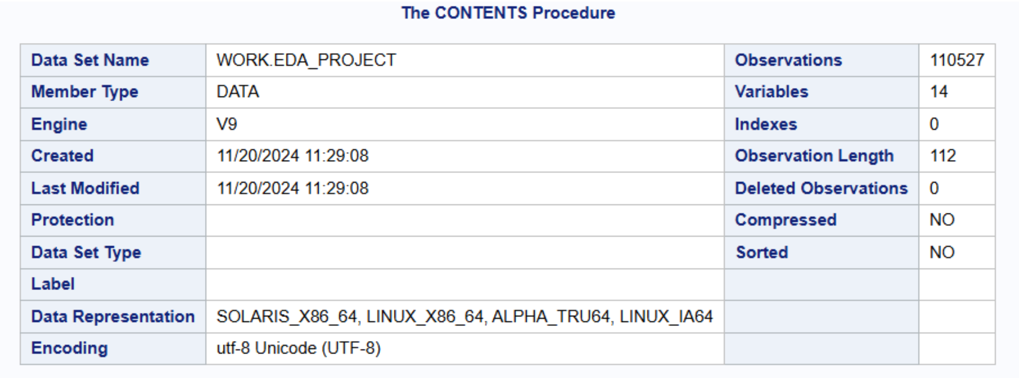
The dataset was now ready for preprocessing and exploratory analysis.

3.2 Data Wrangling

To prepare the data for analysis, the following steps were undertaken:

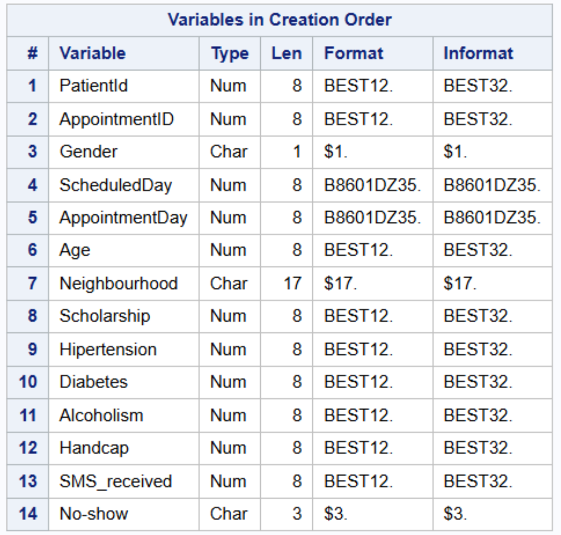
Variable Summary:

- The PROC CONTENTS procedure was used to get a summary of all variables, including data types, lengths, and labels.
- Observations such as variable names, types, and formats were noted.



The CONTENTS Procedure

Data Set Name	WORK.EDA_PROJECT	Observations	110527
Member Type	DATA	Variables	14
Engine	V9	Indexes	0
Created	11/20/2024 11:29:08	Observation Length	112
Last Modified	11/20/2024 11:29:08	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

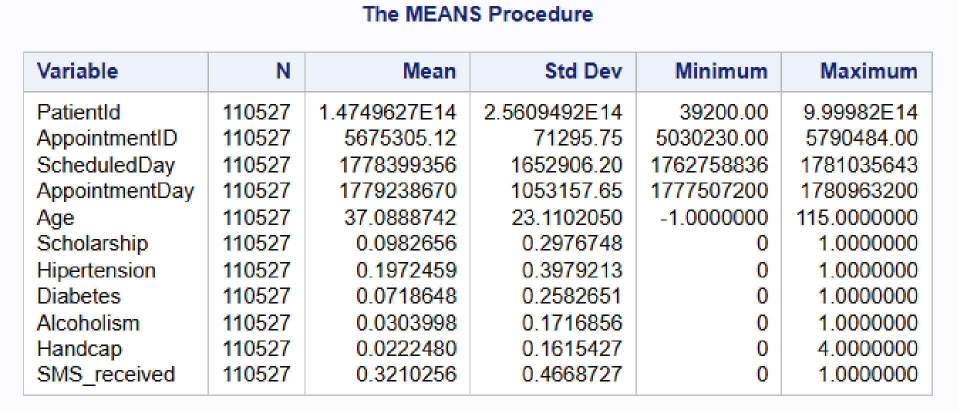


Variables in Creation Order

#	Variable	Type	Len	Format	Informat
1	PatientId	Num	8	BEST12.	BEST32.
2	AppointmentID	Num	8	BEST12.	BEST32.
3	Gender	Char	1	$1.	$1.
4	ScheduledDay	Num	8	B8601DZ35.	B8601DZ35.
5	AppointmentDay	Num	8	B8601DZ35.	B8601DZ35.
6	Age	Num	8	BEST12.	BEST32.
7	Neighbourhood	Char	17	$17.	$17.
8	Scholarship	Num	8	BEST12.	BEST32.
9	Hipertension	Num	8	BEST12.	BEST32.
10	Diabetes	Num	8	BEST12.	BEST32.
11	Alcoholism	Num	8	BEST12.	BEST32.
12	Handcap	Num	8	BEST12.	BEST32.
13	SMS_received	Num	8	BEST12.	BEST32.
14	No-show	Char	3	$3.	$3.

Descriptive Statistics:

- The PROC MEANS procedure was employed to compute basic statistics such as mean, minimum, and maximum values for numerical variables.
- This provided insights into the data distribution and highlighted any inconsistencies, like unexpected negative values.

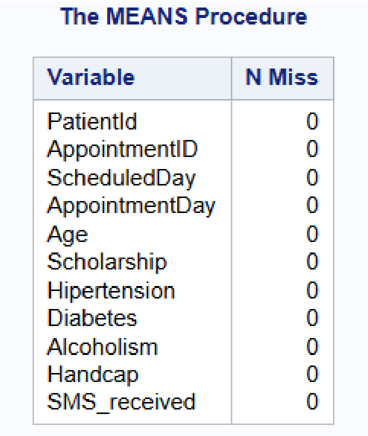


The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
PatientId	110527	1.4749627E14	2.5609492E14	39200.00	9.99982E14
AppointmentID	110527	5675305.12	71295.75	5030230.00	5790484.00
ScheduledDay	110527	1778399356	1652906.20	1762758836	1781035643
AppointmentDay	110527	1779238670	1053157.65	1777507200	1780963200
Age	110527	37.0888742	23.1102050	-1.0000000	115.0000000
Scholarship	110527	0.0982656	0.2976748	0	1.0000000
Hipertension	110527	0.1972459	0.3979213	0	1.0000000
Diabetes	110527	0.0718648	0.2582651	0	1.0000000
Alcoholism	110527	0.0303998	0.1716856	0	1.0000000
Handcap	110527	0.0222480	0.1615427	0	4.0000000
SMS_received	110527	0.3210256	0.4668727	0	1.0000000

Missing Values:

- The dataset was checked for missing values using PROC MEANS and conditional checks.
- Fortunately, no missing values were detected, ensuring completeness of the data.



The MEANS Procedure

Variable	N Miss
PatientId	0
AppointmentID	0
ScheduledDay	0
AppointmentDay	0
Age	0
Scholarship	0
Hipertension	0
Diabetes	0
Alcoholism	0
Handcap	0
SMS_received	0

Through wrangling, the dataset was thoroughly understood, and a plan was formulated for cleaning and transformation.

3.3 Data Cleaning

To ensure data quality and usability, several cleaning steps were implemented:

Variable Renaming:

- The variable hipertension was renamed to Hypertension for consistency.
- The No-show variable was renamed to Show, and its values were inverted and converted to integers (1 for "showed up" and 0 for "no-show").

Date Transformation:

- The ScheduledDay and AppointmentDay columns contained datetime values. Using SAS functions, only the date components were extracted for simplicity and relevance.

New Variable Creation:

- A new column, day_diff, was created to calculate the difference in days between ScheduledDay and AppointmentDay. This variable was essential for analyzing the impact of time gaps on appointment attendance.

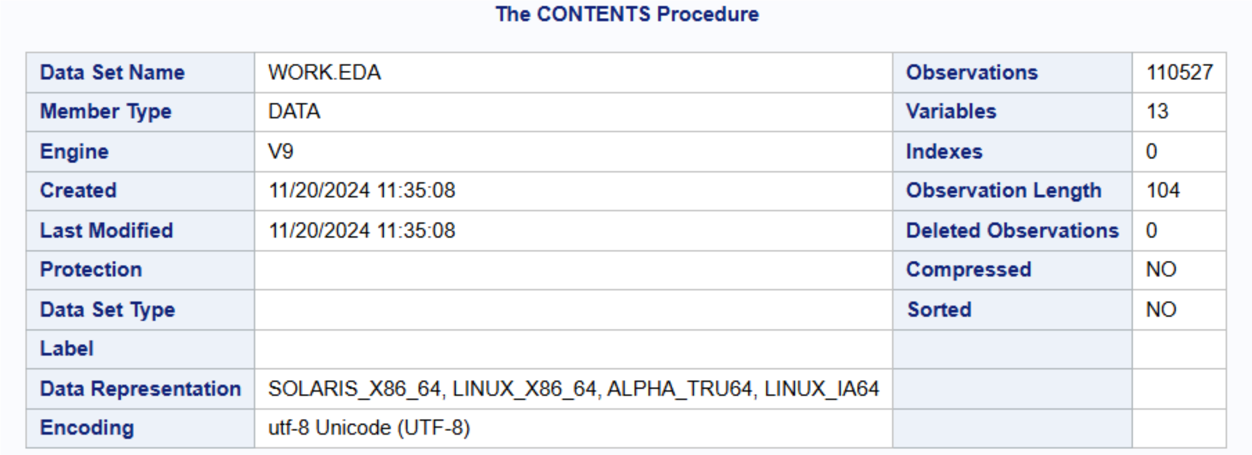
Dropping Irrelevant Columns:

- The PatientId and AppointmentId columns were removed as they did not provide analytical value for the study.

Handling Data Anomalies:

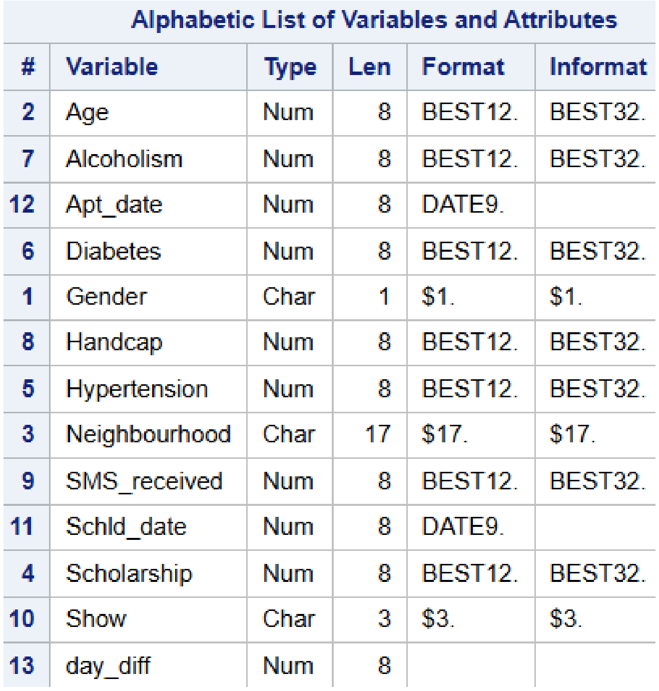
- Negative values in the Age column were interpreted as errors (possibly representing unborn children). These values were excluded from the analysis.
- Any invalid day_diff values (e.g., negative differences) were treated as data inconsistencies and excluded.

After cleaning, the dataset was prepared with appropriate column types, meaningful variables, and a clean structure for analysis. These steps ensured reliable insights during the exploratory data analysis phase.



The CONTENTS Procedure

Data Set Name	WORK.EDA	Observations	110527
Member Type	DATA	Variables	13
Engine	V9	Indexes	0
Created	11/20/2024 11:35:08	Observation Length	104
Last Modified	11/20/2024 11:35:08	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

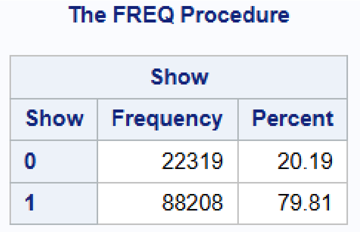


Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat
2	Age	Num	8	BEST12.	BEST32.
7	Alcoholism	Num	8	BEST12.	BEST32.
12	Apt_date	Num	8	DATE9.	
6	Diabetes	Num	8	BEST12.	BEST32.
1	Gender	Char	1	$1.	$1.
8	Handcap	Num	8	BEST12.	BEST32.
5	Hypertension	Num	8	BEST12.	BEST32.
3	Neighbourhood	Char	17	$17.	$17.
9	SMS_received	Num	8	BEST12.	BEST32.
11	Schld_date	Num	8	DATE9.	
4	Scholarship	Num	8	BEST12.	BEST32.
10	Show	Char	3	$3.	$3.
13	day_diff	Num	8		

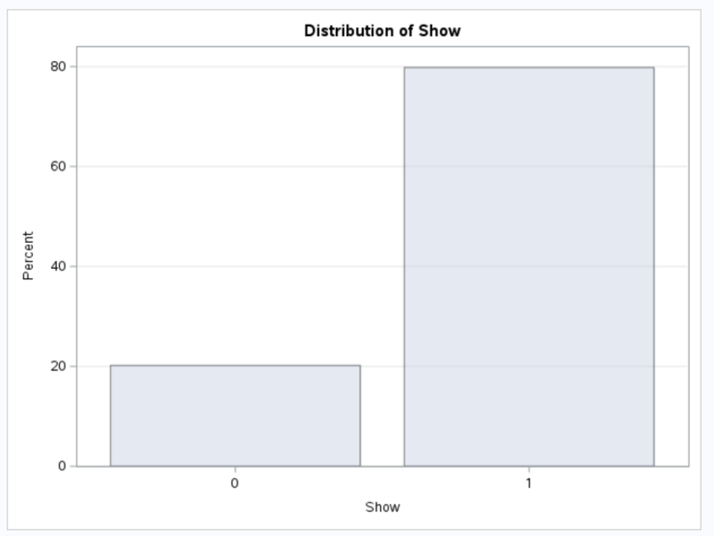
4. Exploratory Data Analysis

4.1 Percentage of No-Shows



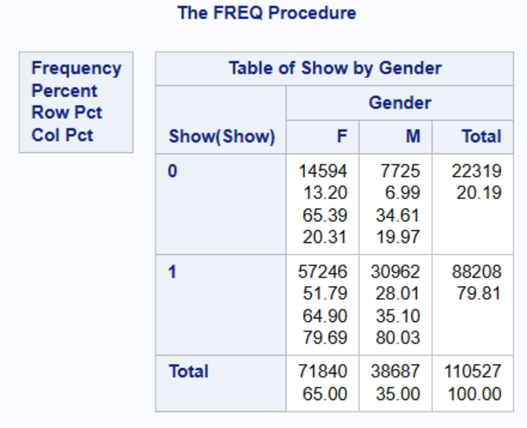
The FREQ Procedure

Show		
Show	**Frequency**	**Percent**
0	22319	20.19
1	88208	79.81



The analysis reveals that 20.19% of patients did not show up for their scheduled medical appointments. This translates to 22,319 missed appointments out of a total of 110,527 scheduled appointments. This percentage highlights a notable gap in patient attendance, which could be explored further for possible interventions.

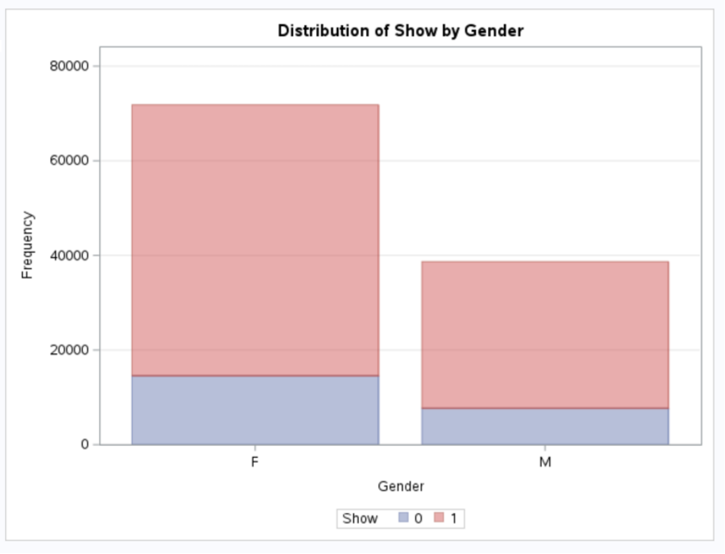
4.2 Gender Analysis



The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of Show by Gender			
	Gender		
Show(Show)	**F**	**M**	**Total**
0	14594 13.20 65.39 20.31	7725 6.99 34.61 19.97	22319 20.19
1	57246 51.79 64.90 79.69	30962 28.01 35.10 80.03	88208 79.81
Total	71840 65.00	38687 35.00	110527 100.00

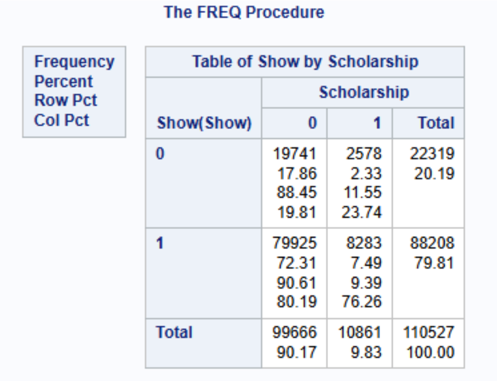


A closer look at the data shows a clear disparity between male and female patients:

- Females who missed appointments: 13.2% of total appointments.
- Males who missed appointments: 6.99% of total appointments.

This analysis indicates that females are nearly twice as likely to miss their appointments compared to males, possibly reflecting gender-specific challenges or responsibilities.

4.3 Effect of Scholarship on Attendance

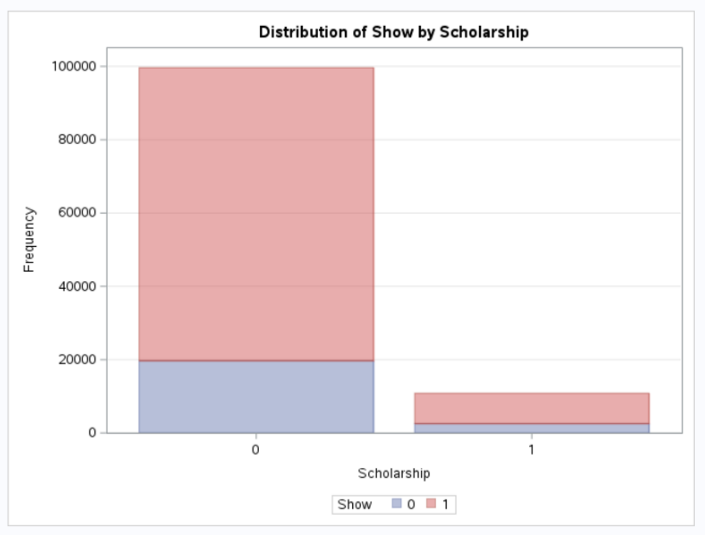


The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of Show by Scholarship

	Scholarship		
Show(Show)	0	1	Total
0	19741 17.86 88.45 19.81	2578 2.33 11.55 23.74	22319 20.19
1	79925 72.31 90.61 80.19	8283 7.49 9.39 76.26	88208 79.81
Total	99666 90.17	10861 9.83	110527 100.00

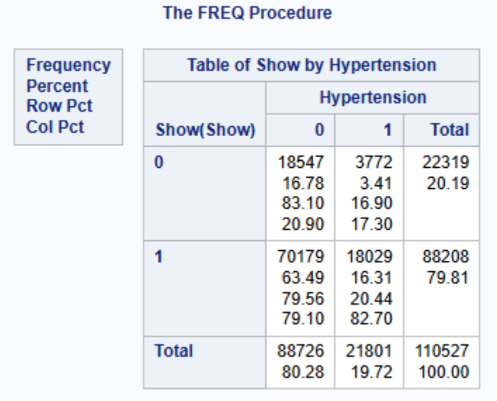


Patients receiving financial aid through the scholarship program are slightly more likely to miss their appointments compared to those who do not.

- Attendance rate without scholarship: 80.19%.
- Attendance rate with scholarship: 76.26%.

This finding suggests that economic support alone may not ensure attendance, and other barriers may need to be addressed.

4.4 Hypertension and Attendance Correlation

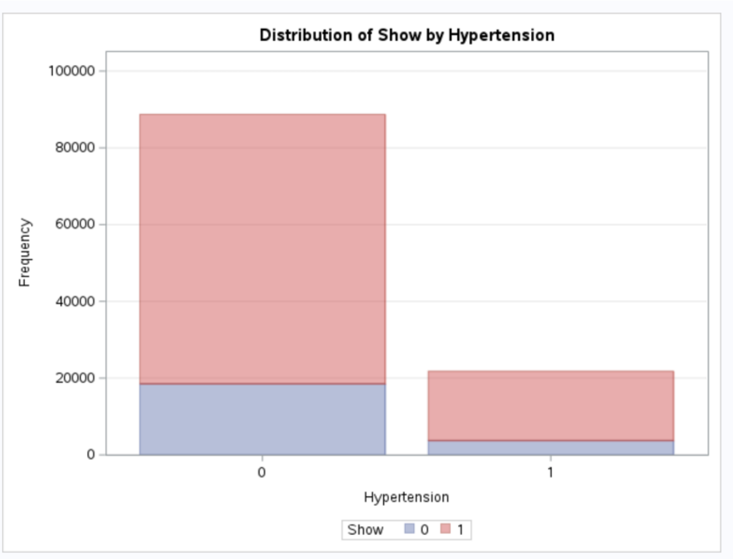


The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of Show by Hypertension

	Hypertension		
Show(Show)	0	1	Total
0	18547 16.78 83.10 20.90	3772 3.41 16.90 17.30	22319 20.19
1	70179 63.49 79.56 79.10	18029 16.31 20.44 82.70	88208 79.81
Total	88726 80.28	21801 19.72	110527 100.00

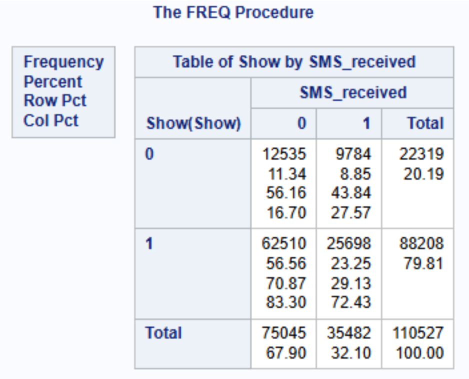


Interestingly, patients with hypertension are more likely to show up for their appointments compared to those without hypertension:

- Attendance rate without hypertension: 79.56%.
- Attendance rate with hypertension: 82.7%.

This suggests that patients with chronic conditions such as hypertension might prioritize their healthcare more consistently.

4.5 Impact of SMS on Attendance

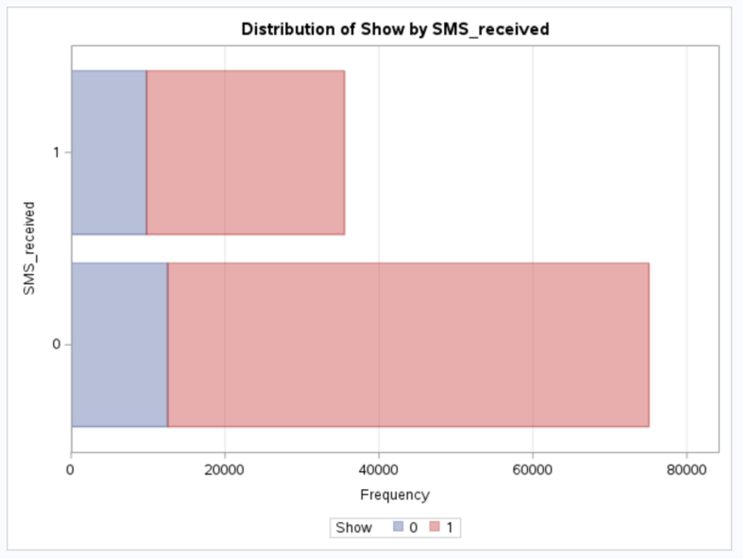


The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of Show by SMS_received

	SMS_received		
Show(Show)	**0**	**1**	**Total**
0	12535 11.34 56.16 16.70	9784 8.85 43.84 27.57	22319 20.19
1	62510 56.56 70.87 83.30	25698 23.25 29.13 72.43	88208 79.81
Total	75045 67.90	35482 32.10	110527 100.00

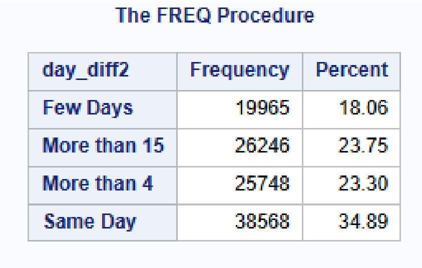


Contrary to expectations, patients who received an SMS reminder were more likely to miss their appointments than those who did not receive an SMS:

- Attendance rate without SMS: 83.3%.
- Attendance rate with SMS: 72.43%.

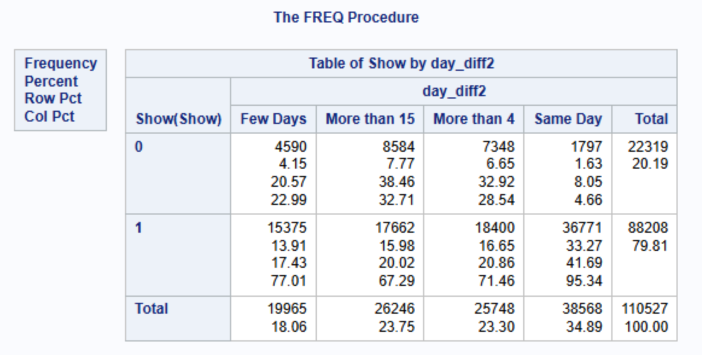
This counterintuitive finding may reflect factors such as over-reliance on reminders or other external influences.

4.6 Time Difference Analysis (Scheduling vs. Appointment)



The FREQ Procedure

day_diff2	Frequency	Percent
Few Days	19965	18.06
More than 15	26246	23.75
More than 4	25748	23.30
Same Day	38568	34.89

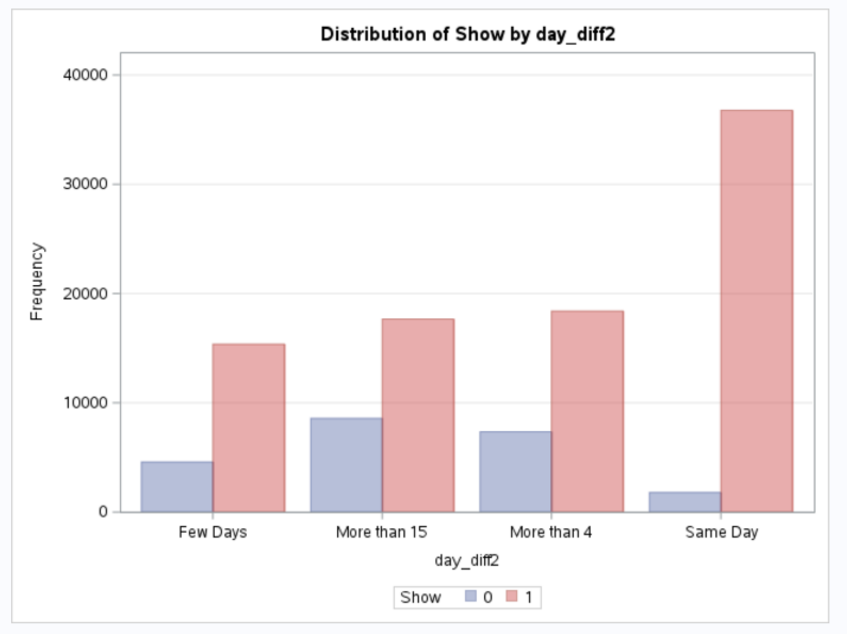


The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of Show by day_diff2

	day_diff2				
Show(Show)	**Few Days**	**More than 15**	**More than 4**	**Same Day**	**Total**
0	4590 4.15 20.57 22.99	8584 7.77 38.46 32.71	7348 6.65 32.92 28.54	1797 1.63 8.05 4.66	22319 20.19
1	15375 13.91 17.43 77.01	17662 15.98 20.02 67.29	18400 16.65 20.86 71.46	36771 33.27 41.69 95.34	88208 79.81
Total	19965 18.06	26246 23.75	25748 23.30	38568 34.89	110527 100.00

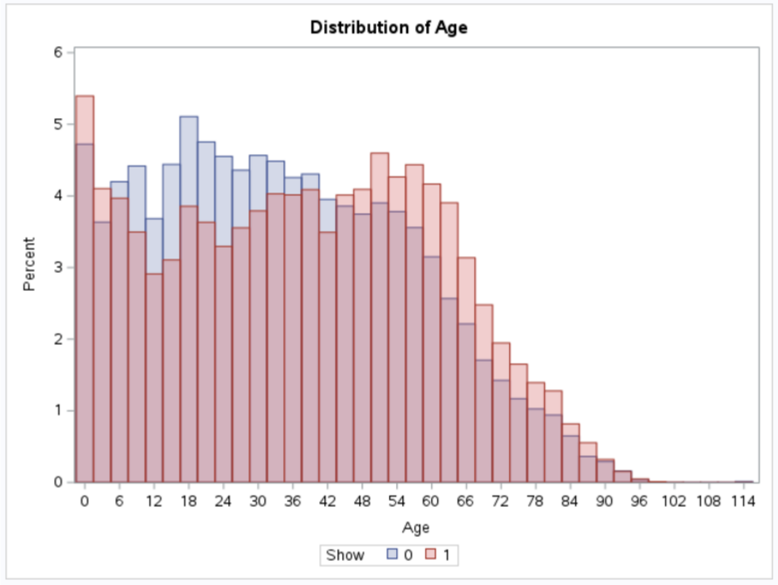


The analysis shows a strong relationship between the time gap between scheduling and the actual appointment day and the likelihood of attendance:

- Same-day appointments: 91.91% attendance.
- Few days' gap (1-3 days): 79.39% attendance.
- 4-15 days gap: 67.14% attendance.
- More than 15 days gap: 61.56% attendance.

This indicates that longer waiting periods correlate with a higher likelihood of no-shows, emphasizing the importance of minimizing delays.

4.7 Age and Attendance Relationship

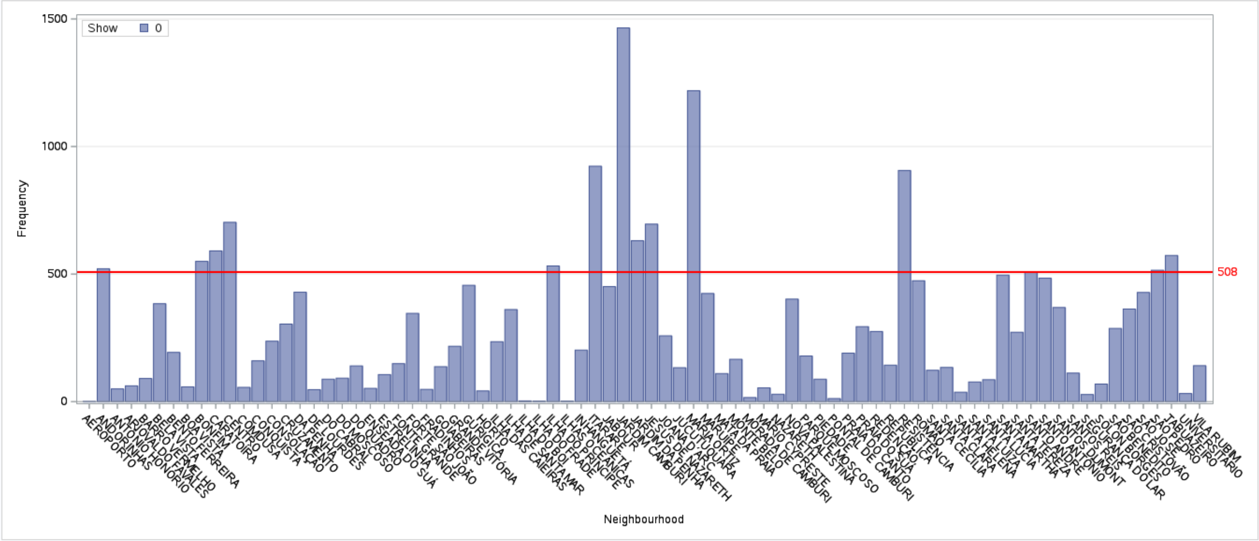


While no direct correlation between age and attendance was observed, specific age groups show higher tendencies to miss appointments:

- Younger patients (6-42 years) are more likely to miss appointments.

This trend could be attributed to lifestyle factors, work commitments, or lesser perceived urgency for healthcare among these age groups.

4.8 Neighborhood Analysis of No-Shows

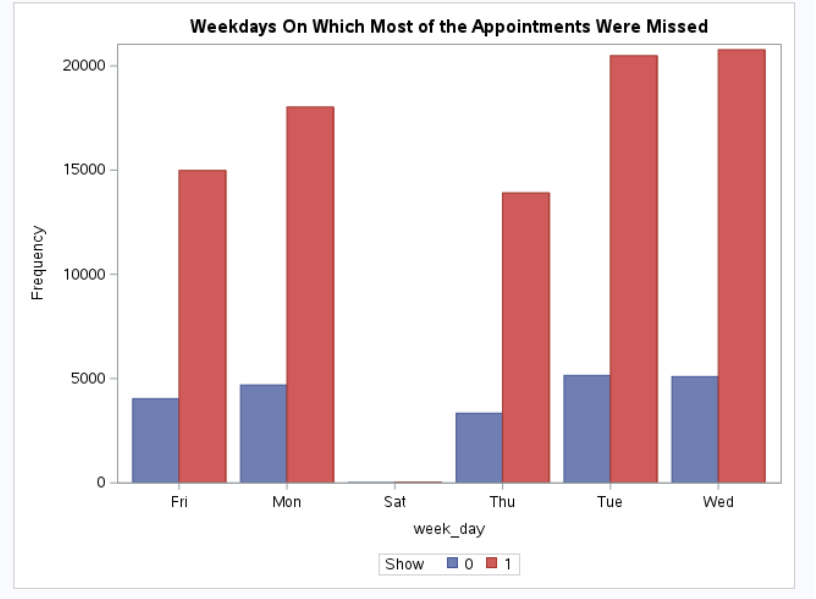


The dataset identifies neighborhoods with the highest number of no-shows:

- Jardim Camburi: 1,465 no-shows.
- Maria Ortiz: 1,219 no-shows.
- Itarare: 923 no-shows.
- Resistencia: 906 no-shows.
- Centro: 703 no-shows.

These insights can guide targeted interventions in specific communities to improve patient attendance.

4.9 Day of the Week Analysis



The day of the week also influences no-show rates:

- Most missed appointments occur on Tuesdays and Wednesdays, which are also the days with the highest number of scheduled appointments.
- This pattern suggests a possible workload imbalance or scheduling inefficiencies that might impact patient turnout.

5. Findings and Observations

5.1 Key Insights

Overall Attendance:

20.19% of patients did not show up for their appointments, highlighting a significant proportion of missed appointments.

Gender Analysis:

Female patients are more likely to miss their appointments compared to male patients, with the percentage of missed appointments for females being nearly double that of males.

(Females: 13.2%, Males: 6.99%)

Scholarship Impact:

Patients with scholarships are slightly more likely to miss their appointments compared to those without scholarships.

SMS Notifications:

Unexpectedly, patients who received SMS notifications were more likely to miss their appointments than those who did not receive any messages.

Time Difference Between Scheduling and Appointment:

The likelihood of missing an appointment increases as the time gap between scheduling and the appointment day grows.

(e.g., Same-day appointments have the lowest no-show rate at 8.09%, while those scheduled more than 15 days in advance have the highest no-show rate at 38.44%.)

Age Group Trends:

Younger patients, particularly those aged 6 to 42 years, are more likely to miss their appointments.

Neighborhood Trends:

Specific neighborhoods, such as Jardim Camburi, Maria Ortiz, and ITARARE, have the highest number of missed appointments. Jardim Camburi alone accounts for 1,465 missed appointments.

Day of the Week:

Tuesdays and Wednesdays have the highest number of missed appointments, but they also see the highest attendance rates.

5.2 Patterns and Anomalies

Patterns Observed:

- Scheduling Gaps: The longer the gap between scheduling and the appointment, the more likely a no-show occurs. This indicates that shorter lead times may encourage better attendance.
- Neighborhood Clusters: Certain neighborhoods consistently have higher no-show rates, which may indicate localized challenges such as accessibility or awareness.

Anomalies:

- SMS Effect: It was expected that sending SMS notifications would improve attendance rates, but the data shows the opposite. Patients who received SMS notifications were more likely to miss their appointments.
- Age Outlier: The age column contained a negative value, which was clarified as representing an unborn baby (a pregnant woman), an unconventional representation in the dataset.
- Negative Time Difference: Some scheduling dates were recorded as being after the appointment date, suggesting potential data entry errors.

6. Conclusion

The analysis identified key factors influencing patient attendance at medical appointments, highlighting the following:

- Gender: Female patients were more likely to miss their appointments than male patients.
- Scholarship status: Patients with scholarships showed a higher likelihood of missing their appointments.
- SMS reminders: Surprisingly, patients who received SMS reminders were more likely to miss their appointments.
- Scheduling vs. appointment day gap: Longer gaps between scheduling and the appointment date were associated with higher no-show rates.
- Age: No significant correlation was found between age and attendance, although younger patients (6-42 years) were more likely to miss appointments.
- Neighborhoods: Certain neighborhoods, such as Jardim Camburi and Maria Ortiz, exhibited the highest no-show rates.
- Weekdays: Tuesdays and Wednesdays had the highest number of missed appointments.

Practical Implications:

- Tailored interventions, such as improved reminder strategies, could help reduce no-show rates, particularly for patients with scholarships or in high-risk neighborhoods.
- Reevaluating scheduling practices and reducing the time gap between scheduling and appointments may help lower the number of no-shows.
- Exploring alternative methods for reminding patients, particularly those who have shown a tendency to miss appointments despite SMS reminders, could improve attendance.
- A deeper investigation into the unexpected trends, such as SMS reminders increasing no-show rates, could help optimize patient engagement strategies.

Impact:

- The findings provide actionable insights for healthcare providers to reduce no-shows, improve patient attendance, and optimize resource utilization.
- Future research should focus on exploring the underlying reasons for no-shows and testing targeted interventions to improve appointment adherence.

7. Challenges and Limitations

7.1 Data Limitations

- Missing Appointment Time: Lack of exact appointment times limits analysis of time-of-day effects on attendance.
- No Holiday Data: Missing information on holidays prevents understanding of their impact on no-shows.
- Negative Age Values: Negative values represent unborn children, which could skew results in age-based analysis.
- Negative Day Difference: Incorrect data showing negative values in scheduling gaps impacts time-related findings.
- Lack of Appointment Type Details: The dataset does not differentiate between various medical specialties, limiting depth.

7.2 Analytical Shortcomings

- Correlation vs. Causality: The analysis shows correlations but does not establish cause-and-effect relationships.
- Age and Attendance: No strong relationship was found between age and attendance despite initial expectations.
- SMS and Attendance: Patients who received SMS reminders were more likely to miss appointments, a counterintuitive result.
- Geographical Factors: Neighborhood data was too general to capture detailed regional differences affecting attendance.
- Data Scope: The dataset doesn't capture other influencing factors (e.g., personal emergencies, weather) that could affect attendance.

References

Dataset Source: The dataset used for this analysis was sourced from Kaggle: Joniarroba's "No-Show Appointments" dataset, available at Kaggle No-Show Appointments.

Online Learning Resource: The original project concept and Python-based analysis were based on the "Medical Appointment Data Analysis" course from Cognitive Class. The course can be accessed at Cognitive Class Course.

SAS Documentation: SAS Institute. (2021). SAS Knowledge Base. Retrieved from SAS Knowledge Base.

SAS Blog on EDA: SAS Institute. (2021). "Understanding Your Data: Visual Exploratory Data Analysis". Retrieved from SAS Blog - Visual EDA.