**EDA on Healthcare Appointment Data Using Python (**[**LINK**](https://youtu.be/kJdta28q7vg?si=P8K2Jirw0xLVd_tM)**)**

🎙️ Interview Explanation – EDA on Healthcare Data

"Sure, I'd love to walk you through a project I worked on recently, which involved exploratory data analysis (EDA) on a healthcare appointment dataset.

The main goal of this project was to analyze the factors that influence whether patients show up for their medical appointments or not. It was a real-world dataset with over 110,000 patient records, and instead of building a predictive model, I focused entirely on descriptive analytics to extract meaningful patterns from the data.

📂 Dataset and Business Context

The dataset included various attributes like:

* Patient demographics (gender, age, neighbourhood)
* Medical indicators (hypertension, diabetes, scholarship, handicap, etc.)
* Time-related data like scheduled day and appointment day
* A column indicating whether the patient actually showed up (no-show)

The No-show column was our target variable, where “No” means the patient showed up, and “Yes” means they didn’t.

🛠️ Steps I Followed

I started with data cleaning:

* I removed irrelevant columns like PatientID and AppointmentID since they’re just identifiers and don’t contribute to insights.
* I also fixed some typos in column names — for example, Handcap was renamed to Handicap for clarity.

Next, I handled the datetime columns:

* Converted the ScheduledDay and AppointmentDay into proper datetime formats.
* From these, I extracted the day of the week to analyze if appointment behavior varied across weekdays vs weekends.

🔢 Feature Engineering and Encoding

I also binned the age column into groups like 0–20, 21–40, and so on. This helped in visualizing trends across age ranges rather than individual ages, which were too granular.

For categorical variables like gender and age group, I used one-hot encoding via get\_dummies to prepare them for correlation analysis and visualizations.

📊 Analysis and Visualizations

Once the data was ready, I moved on to univariate and bivariate analysis:

* I visualized the distribution of the target variable and noticed that the data was highly imbalanced: about 80% of patients showed up and 20% didn’t.
* I plotted bar charts and heatmaps to understand how different factors relate to the no-show behavior.

📈 Key Insights

Some of the key findings were:

1. Female patients scheduled more appointments than male patients.
2. The show rate was consistent across all neighborhoods — about 80% showed up, regardless of location.
3. Patients with hypertension and diabetes were more likely to attend their appointments — their show rate was as high as 85%.
4. Surprisingly, patients who received SMS reminders were less likely to show up than those who didn’t — which went against our expectations.
5. No appointments were scheduled on Sundays, and very few were on Saturdays. Most appointments happened between Monday and Wednesday.
6. The youngest age groups — particularly Age 0 and Age 1 — had the highest show rate, around 80%.

📌 Business Impact and What I’d Do Next

From a business perspective, this analysis can help optimize appointment scheduling — for example, focus reminders differently, or target patients with chronic conditions more carefully.  
If I were to take this further, I’d build a classification model using these features to predict no-shows, or create a Power BI dashboard for hospital administrators to make informed scheduling decisions in real time.

✅ Conclusion

Overall, this project gave me hands-on experience with real-world messy data, and helped me improve my skills in data cleaning, feature engineering, and uncovering actionable insights using Python, pandas, and seaborn. It’s also a great example of how EDA alone can add a lot of value before even touching machine learning.

Let me know if you'd like a short version too — for quick answers during rapid-fire rounds — or if you'd like help preparing follow-up questions interviewers might ask about this project.

**QUESTION on Project**

**🔍 EDA & Data Understanding**

1. **Why did you drop PatientID and AppointmentID? Could they contain useful information?**

(They’re auto-generated identifiers; they don’t carry predictive value and may add noise.)

1. **Why did you bin the Age column? What if you lost information in the process?**

(To simplify the analysis and avoid cluttered visualizations; but yes, it’s a trade-off.)

1. **How did you handle missing values? What would you do if there were missing values in age or medical history?**

(You can mention using mean/median or model-based imputation depending on the situation.)

**📅 Time-based Features**

1. **Why did you extract the day of the week from the appointment date?**

(To analyze weekly trends — are no-shows more common on certain days?)

1. **Did you consider the gap between ScheduledDay and AppointmentDay as a feature?**

(This is a great follow-up. If you didn’t, you can say you would add that in future — longer gaps may lead to no-shows.)

**📊 Analysis and Insights**

1. **Why do you think patients who received SMS reminders had *lower* show-up rates?**

(You could hypothesize: maybe SMS was sent to less reliable patients, or it wasn’t effective.)

1. **Did you perform any hypothesis testing or statistical validation of your insights?**

(Even if not, mention you could use t-tests or chi-square tests to validate assumptions.)

1. **Were there any surprising or counterintuitive findings?**

(Mention SMS example, or high show rate for babies.)

**📈 Visualization & Communication**

1. **What types of plots did you use and why?**

(Bar plots for categorical distributions, heatmaps for correlation, etc.)

1. **How would you present this to a non-technical stakeholder?**

(Use dashboards, focus on actionable insights, avoid technical jargon.)

**🤖 Next Steps / ML Thinking**

1. **What kind of model would you build to predict no-shows?**

(Logistic Regression, Decision Trees, Random Forests – handle categorical variables well.)

1. **How would you deal with the imbalanced target variable (80% show, 20% no-show)?**

(Use techniques like oversampling, undersampling, SMOTE, or class weights.)

1. **How would you evaluate the performance of such a model?**

(Precision, recall, F1-score – especially focusing on recall for the minority class.)

**🧠 Reflection & Real-world Thinking**

1. **If you had more data, what features would you like to include?**

(Appointment reason, past no-show history, time of day, patient distance from hospital.)

1. **What would you do differently if you started this project again?**

(Possibly create new features like days between scheduling and appointment, or time-of-day slots.)

**📚 Bonus: Behavioral/Contextual**

1. **What was the biggest challenge you faced during this project?**
2. **How did you validate the insights you derived?**
3. **How would you turn this into a dashboard for hospital management?**

**🎓 Tips for Answering**

* Be **honest** if you didn’t do something — but say what you *would* do next.
* Keep answers **short, structured, and confident**.
* Whenever possible, link your answers to **business impact**.

Would you like me to prepare a **mock Q&A script** with ideal answers for some of these questions? Or a **cheat sheet** for last-minute prep?