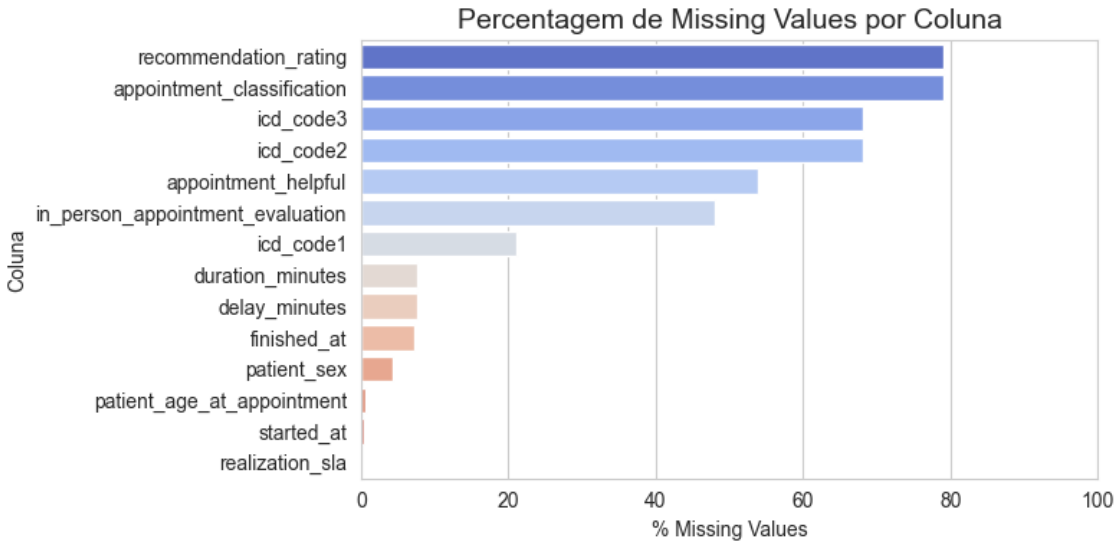
01_EDA - Projeto Knok

Objetivo: Analisar os dados de teleconsultas, identificar missing values, outliers e relações entre variáveis

EDA: Inspeção Inicial

Percentagem de Valores Nulos



Algumas variáveis apresentam um elevado número de missing values

- recommendation_rating (~79% missing)
- appointment_classification (~79% missing)
- icd_code2 / icd_code3 (~68% missing)
- appointment_helpful (~54% missing)

Estas variáveis não serão consideradas na análise principal devido ao risco de bias e à reduzida cobertura de dados

Identificação de variáveis de interesse

Começamos por identificar as variáveis contínuas de interesse:

Principais variáveis contínuas de interesse para detecção de outliers

- realization_sla → tempo desde a marcação até ao início
- duration_minutes → duração da consulta
- delay_minutes → tempo de atraso

Secundárias

- recommendation_rating → maioritariamente de 0 a 10, pode verificar-se a existência de anomalias, mas geralmente menos extremas
- patient_age_at_appointment → apenas alguns erros de digitação, normalmente seguro ignorar para boxplots de outliers.

Quando comparamos os máximos e mínimos com o valor médio das estatísticas descritiva, 3 variáveis continuas destacam-se pelos seus valores extremos:

	min	max	mean
realization_sla	-60.88	100522.77	222.40
duration_minutes	0.02	51368.58	10.67
delay_minutes	-5560.65	5967.67	4.95

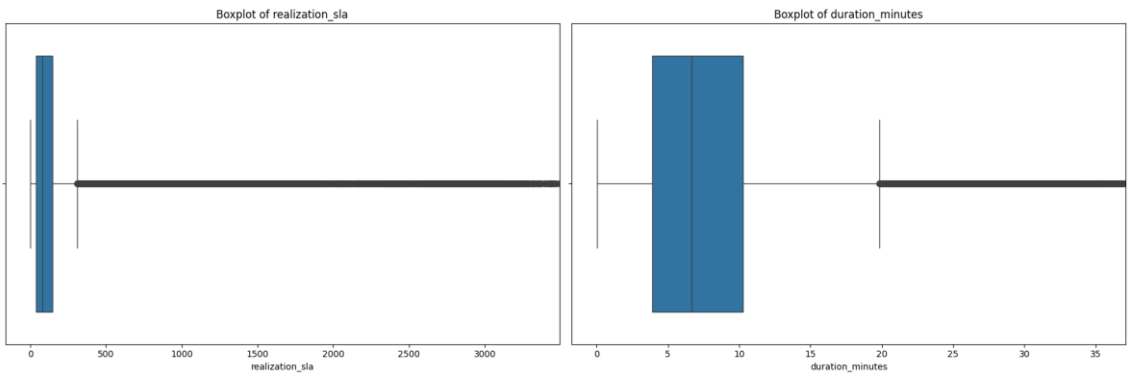
Os valores extremos destas variáveis indicam a presença de outilers. Adicionalmente, valores negativos não têm qualquer interpretação, o que significa que podem estar associados a erros do sistema.

- A variável delay_minutes contém valores muito altos e baixos, mas devido à elevada presença de valores negativos (34.15%) esta variável pode não apresentar grande interpretabilidade. Logo, vai ser descartada desta análise

Pela mesma lógica, valores negativos de duration_minutes e realization_sla foram removidos.

Vamos então olhar para a distribuição destas variáveis para confirmar ou rejeitar estas suspeitas.

Identificação de outliers



Todos estes pontos depois do terceiro quartil (Q3) nos boxplots indicam que existem muitos valores acima do limite superior esperado, ou seja, outliers positivos.

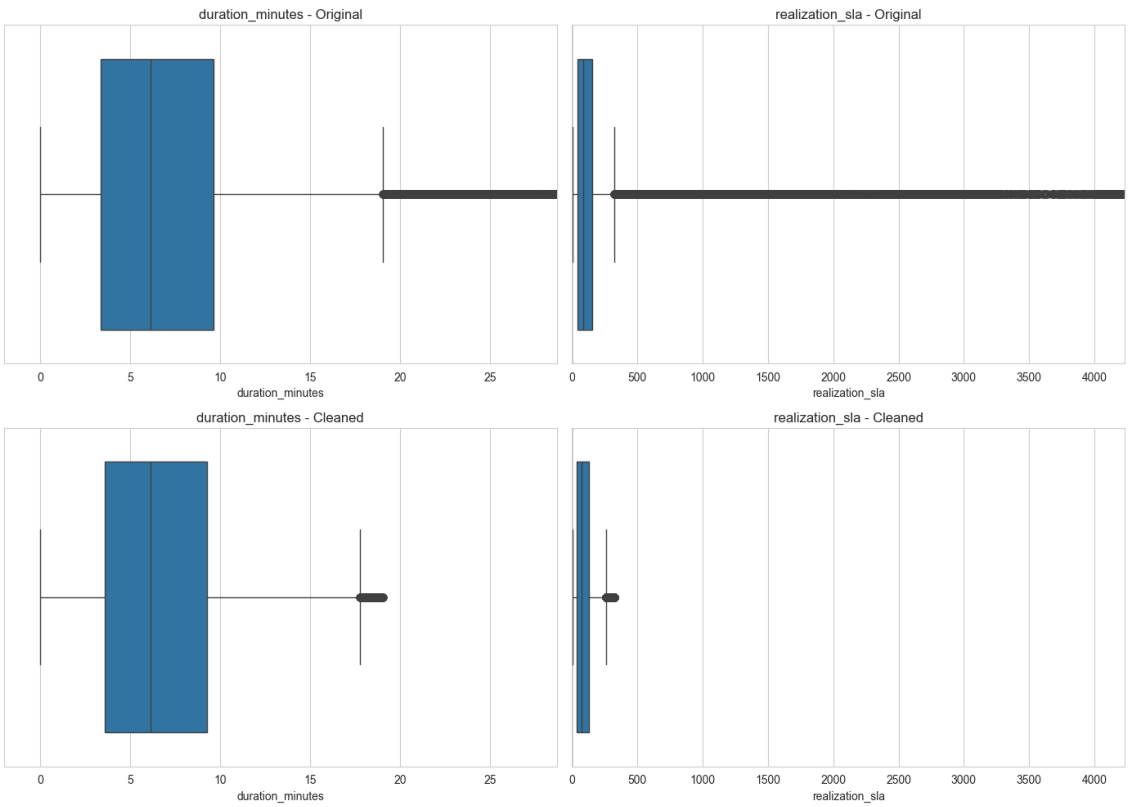
	Variable	Lower Bound	Upper Bound	Outliers (count)	Outliers (%)
0	realization_sla	0	310.80	17240	8.53
1	duration_minutes	0	19.82	7802	3.86

Podemos concluir o seguinte:

- duration_minutes → apresenta valores absolutamente extremos que são claramente impossíveis para uma teleconsulta. Isto indica um problema evidente de outliers.
- realization_sla → apresenta valores bastante elevados que embora não tão extremos quanto duration_minutes, podem distorcer a análise e também devem ser tratados.

Por isso, para aplicar o método de Tukey, faz mais sentido focar nas duas primeiras: duration_minutes e realization_sla.

Método de Turkey

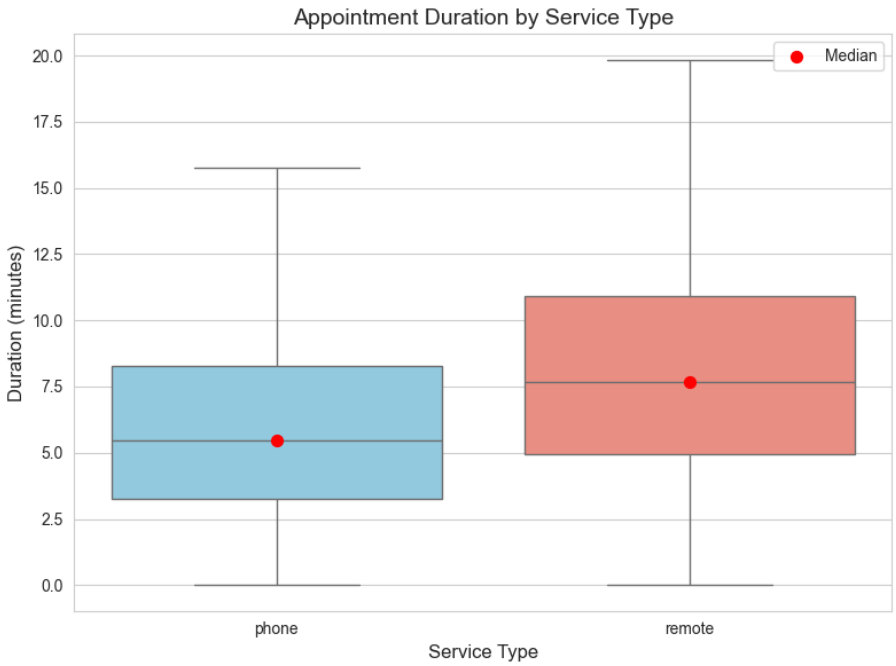


Uma vez removidos os outliers, a integridade do dataset mantém-se uma vez que apenas as removemos 3.30% dos pontos de duration_minutes e 9.65% dos ponstos de realization_sla.

Explorar Associações

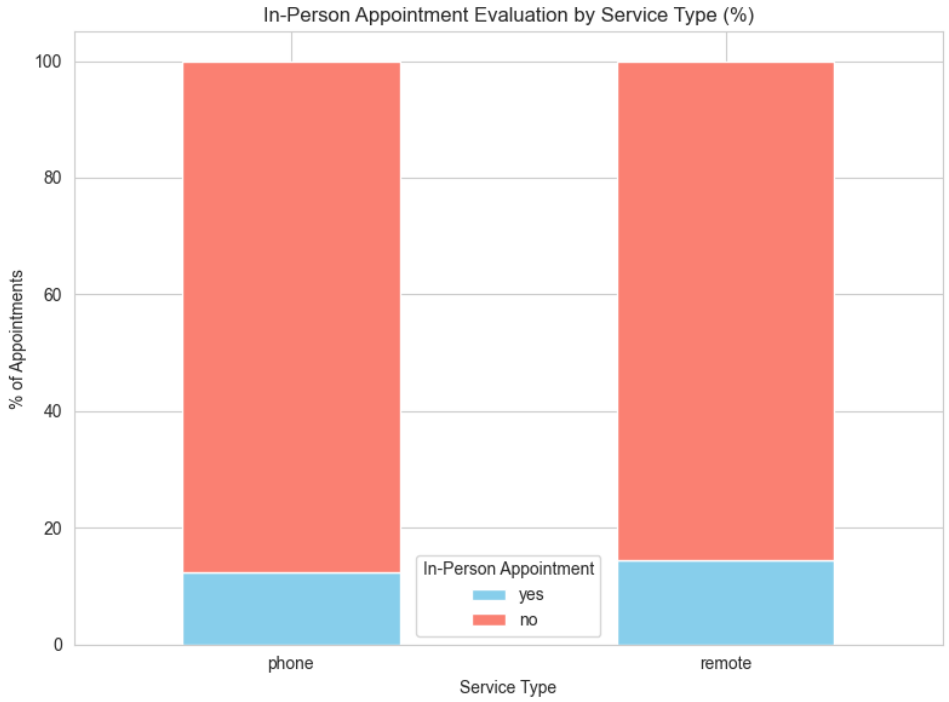
	Variable	Test	p-value	Effect Size	Interpretation
0	duration_minutes	Mann-Whitney U	0.000000e+00	-2.200000	Difference in median
1	in_person_appointment_evaluation	Chi2	6.835912e-274	0.114465	Effect size Cramer's V
2	recurrence_7days	Chi2	3.220233e-267	0.082841	Effect size Cramer's V

Servisse_type vs duration_minutes (duração da consulta)



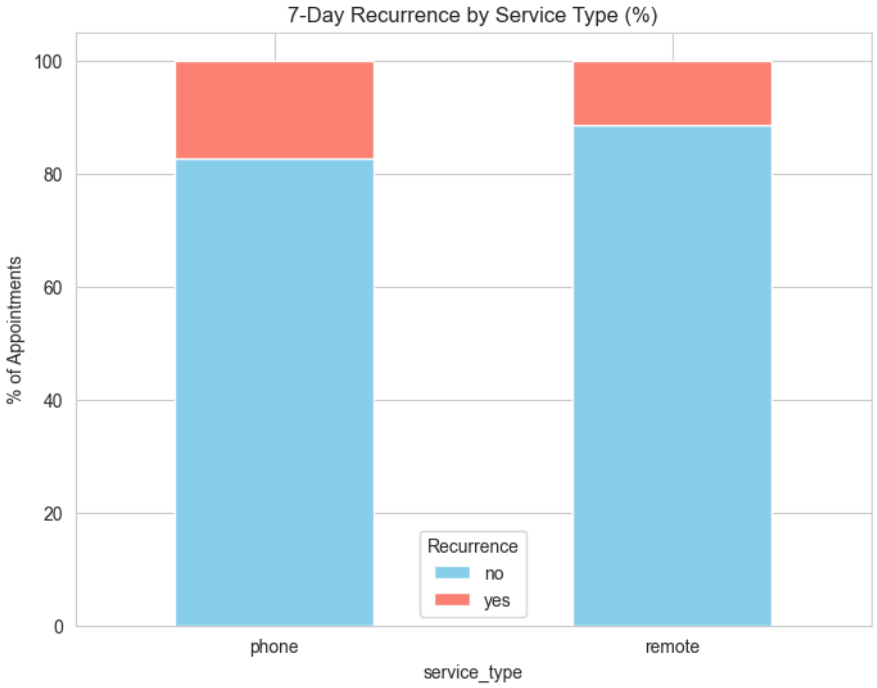
- Consultas mais longas ou mais curtas podem indicar eficiência, complexidade do atendimento ou problemas operacionais.
- p-value: 0.000 , indica que existe diferença estatisticamente significativa entre a duração das consultas dos dois tipos de serviço.
- Apesar da diferença ser estatisticamente significativa, o efeito é relativamente pequeno em termos práticos.

Servisse_type vs in_person_appointment_evaluation (consulta presencial resultante)



- Permite avaliar o impacto do tipo de serviço na necessidade de seguimento presencial, informação relevante para gestão clínica e operacional.
- p-value: ~6.8e-274, diferença altamente significativa entre tipos de serviço relativamente à probabilidade de resultar numa consulta presencial.
- Cramer's V: 0.114, efeito pequeno a moderado.
- Interpretação: O tipo de serviço influencia ligeiramente a probabilidade de necessidade de consulta presencial, mas o efeito não é muito forte.

Servisse_type vs recurrence_7days (recorrência em 7 dias)



- Permite avaliar se um tipo de serviço específico está associado a mais consultas repetidas, que podem sinalizar necessidade de melhoria no processo.
- p-value: ~3.2e-267, diferença altamente significativa na recorrência de serviços em 7 dias entre os tipos de serviço.
- Cramer's V: 0.083, efeito pequeno.
- Interpretação: Embora a associação seja estatisticamente significativa, o efeito prático é pequeno. A recorrência em 7 dias não é fortemente dependente do tipo de serviço.