

Pima Diabetes Data Optimization

ROLE

Lead Methodologist & Data Processor

TOOLS

Microsoft Excel & JASP

01 The Challenge

THE CONSTRAINT

To satisfy the rigorous standards of our IMRAD research design, the data needed to meet strict assumptions for **Normality and Homogeneity**.

THE PROBLEM

The dataset contained massive gaps—specifically **48.7% missingness in Insulin**—and significant outliers that threatened the validity of our group's findings.

TECHNICAL STACK UTILIZED

Statistical Methods

Median Imputation, IQR Outlier Detection, Z-Score Standardization

Software

Microsoft Excel (Analysis Toolpak), JASP (Statistical Computing)

02 Methodology: Audit & Repair

A THE AUDIT CONTEXT

Initial audit identified severe gaps in Insulin and Skin Thickness. This step was crucial to calculate the risk before applying imputation strategies.

Variable	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Pregnancies	Diabetes Pedigree Function	Age
Missing Values (%)	0.65	4.56	29.56	48.70	1.43	0	0	0
Imputed Method	Median	Median	Median	Median	Median	N/A	N/A	N/A

B DATA INTEGRITY (IMPUTATION)

“Successfully salvaged 48.7% of the dataset (Insulin Records) using Median Imputation, preventing massive data loss.”

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	33.6	0.627	50	1	1
1	85	66	29	26.6	0.351	31	0	0
8	183	64	23.3	0.672	32	1	1	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	25.6	0.201	30	0	0	0
3	78	50	32	88	31	0.248	26	1
10	115	72	35.3	0.134	29	0	0	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	29	125	32	0.232	54	1
4	110	92	37.6	0.191	30	0	0	0
10	168	74	38	0.537	34	1	1	1
10	139	80	27.1	1.441	57	0	0	0
1	189	60	23	846	30.1	0.398	59	1
5	166	72	19	175	25.8	0.587	51	1
7	100	72	30	0.484	32	1	1	1
0	118	84	47	230	45.8	0.551	31	1
7	107	74	29.6	0.254	31	1	1	1
1	103	30	38	83	43.3	0.183	33	0
1	115	70	30	96	34.6	0.529	32	1
3	126	88	41	235	39.3	0.704	27	0
8	99	84	35.4	0.388	50	0	0	0
7	196	90	39.8	0.451	41	1	1	1
9	119	80	35	29	0.263	29	1	1
11	143	94	33	146	36.6	0.254	51	1
10	125	70	26	115	31.1	0.205	41	1

BEFORE

AFTER

... Methodology (Continued)

C STATISTICAL READINESS (Z-SCORES)

Standardized variables into Z-Scores (Mean=0, SD = 1). This transformation was mathematically necessary to satisfy assumptions for the group's Multivariate Analysis.

Preg_Capped	Gluc_Capped	BP_Capped	ST_Capped	Insulin_Capped	BMI_Capped	DPF_Capped	Age_Capped	Outcome	Outcome_Z_Preg	Z_Gluc	Z_BP	Z_ST	Z_Insulin	Z_BMI	Z_DPF	Z_Age	
6	148	72	35	71	33.6	0.627	50	1	1	0.647148674	0.885078102	0.541041178	0.888920509	0.059827525	0.355704006	0.599261958	1.445690962
1	86	66	29	71	26.6	0.351	31	0	0	-0.848969981	-0.262239216	0.335901175	0.28913753	0.059827525	-0.42017285	-0.378098994	-0.189304002
8	183	64	27	71	23.3	0.672	32	1	1	1.245597535	1.522476611	0.267521173	0.062543204	0.059827525	-0.785943393	0.746983653	-0.103251635
1	89	66	23	83	28.1	0.167	21	0	0	-0.848969981	-0.189393672	0.335901175	-0.350645448	1.407921584	-0.253913512	-1.023006295	-1.049827666
0	137	4	35	83	43.1	1.199	33	1	1	-1.148193912	0.884752856	-1.783878863	0.888920509	1.407921584	1.408679865	2.594082482	-0.017199269
5	116	74	27	71	25.6	0.201	30	0	0	0.347925743	0.30231375	0.609421179	0.062543204	0.059827525	-0.531012408	-0.903638642	-0.275356368
3	78	5	32	83	31	0.248	26	1	1	-0.250522119	-0.389718918	-1.749688863	0.57902902	1.407921584	0.067521208	-0.739106885	-0.619656834
10	115	66	27	71	35.3	0.134	29	0	0	1.844045397	0.284102364	0.335901175	0.062543204	0.059827525	0.544131309	-1.138669018	-0.361408735
2	197	7	45	83	10.05	0.158	53	1	1	-0.54974805	1.777436015	-1.681308861	1.92189214	1.407921584	-2.254567542	-1.054550674	1.703848081
8	125	96	27	71	32	0.232	54	1	1	1.245597535	0.466216224	1.361601193	0.062543204	0.059827525	0.178360766	-0.795185781	1.789900428
4	11	92	27	71	37.6	0.191	30	0	0	0.048701812	-1.608881779	1.22484119	0.062543204	0.059827525	0.799062284	-0.938887951	-0.275356368
10	168	74	27	71	39	0.537	34	1	1	1.844045397	1.249305821	0.609421179	0.062543204	0.059827525	0.843398117	0.27381817	0.068853098
10	139	8	27	71	27.1	1.199	67	0	0	1.844045397	0.721175628	-1.647118861	0.062543204	0.059827525	-0.36475307	2.594082482	2.048057527
1	189	6	23	83	10.05	0.398	69	1	1	-0.848969981	1.631744927	-1.715498862	-0.350645448	1.407921584	-2.254567542	-0.213367237	2.22016226
5	166	72	19	83	25.8	0.587	51	1	1	0.347925743	1.212883049	0.541041178	-0.7638341	1.407921584	-0.508844496	0.449064719	1.531743328
7	1	66	27	71	10.05	0.484	32	1	1	0.946373605	-1.791995639	0.335901175	0.062543204	0.059827525	-2.254567542	0.09056827	-0.103251635
0	118	84	47	57	45.8	0.551	31	1	1	-1.148193912	0.336736522	0.951321166	2.128466466	-1.512848878	1.707946673	0.322867203	-0.189304002
7	17	74	27	71	29.6	0.254	31	1	1	0.946373605	-1.500613463	0.809421179	0.062543204	0.059827525	-0.087654174	-0.718077299	-0.189304002
1	13	3	38	83	43.3	0.183	33	0	0	-0.848969981	-1.573459007	-1.810068864	1.198011998	1.407921584	1.430847777	-0.969927399	-0.017199269
1	115	7	7	83	34.6	0.529	32	1	1	-0.848969981	0.284102364	-1.681308861	-2.003400058	1.407921584	0.466543618	0.245778722	-0.103251635
3	126	88	41	83	39.3	0.704	27	0	0	-0.250522119	0.48442761	0.08081168	1.509703487	1.407921584	0.987489543	0.859141644	-0.533513468
8	99	84	27	71	35.4	0.398	50	0	0	1.245597535	-0.007279812	0.951321166	0.062543204	0.059827525	0.555215265	-0.248416547	1.445690962
7	196	9	27	71	39.8	0.451	41	1	1	0.946373605	1.759224629	-1.61282886	0.062543204	0.059827525	1.042809322	-0.027605895	0.671219683
9	119	8	35	71	29	0.283	29	1	1	1.544821466	0.356947908	-1.647118861	0.888920509	0.059827525	-0.154157909	-0.686532821	-0.361408735
11	143	84	33	83	36.6	0.254	51	1	1	2.143269328	0.794021172	1.283221192	0.682326183	1.407921584	0.688222735	-0.718077299	1.531743328
10	125	7	28	83	31.1	0.205	41	1	1	1.844045397	0.466216224	-1.681308861	-0.040753959	1.407921584	0.078605163	-0.689918918	0.671219683
7	147	76	27	71	39.4	0.257	43	1	1								
1	97	66	15	57	23.7	0.482	22	0	0								

BEFORE

AFTER

BEFORE

AFTER

03 Research Impact



Data Integrity Preserved

Prevented the deletion of nearly half the sample size by successfully imputing 100% of missing **Insulin** records.



Statistical Robustness

The cleaned data set satisfied assumptions for the team's JASP analysis, leading to higher confidence intervals in our findings.