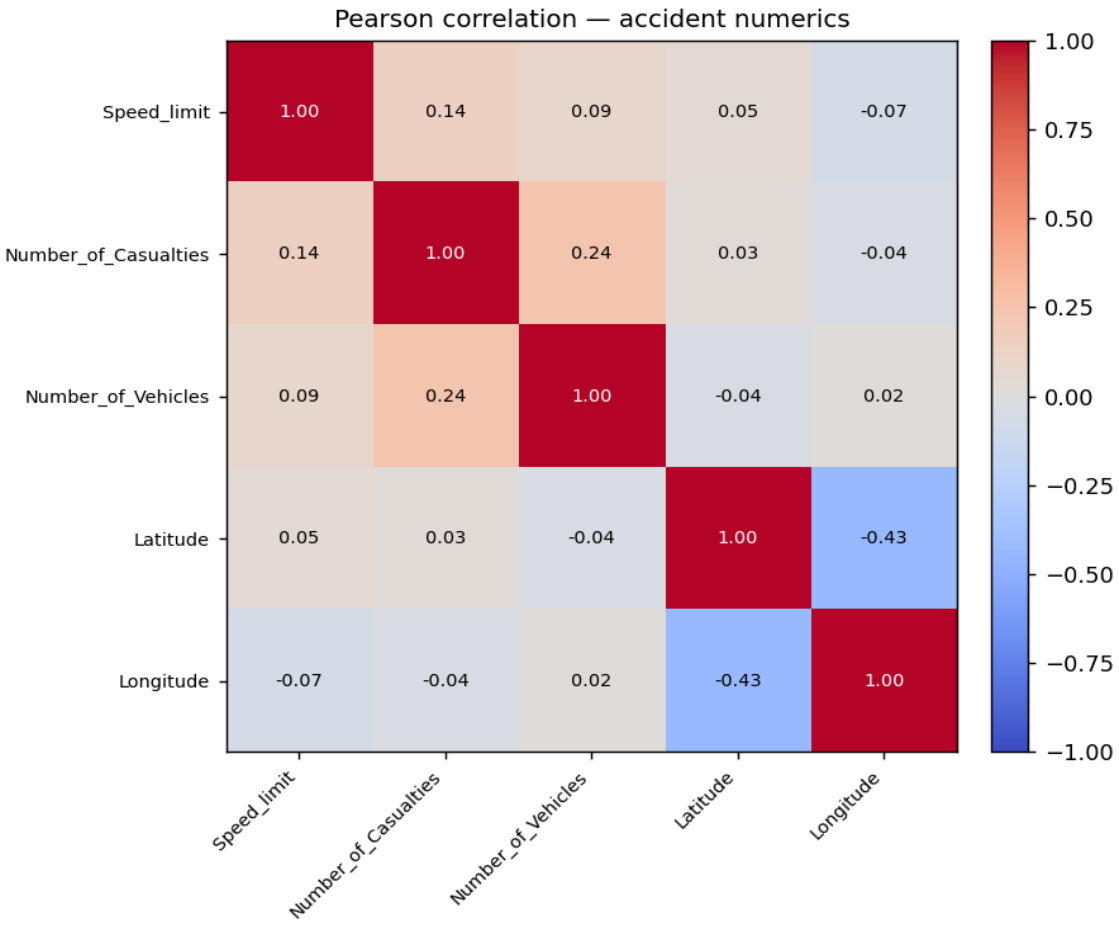
# UK Road Accidents — Data Validation Report

**Generated:** 2026-05-02T05:35:17.191499

## Total Issues Found: 18

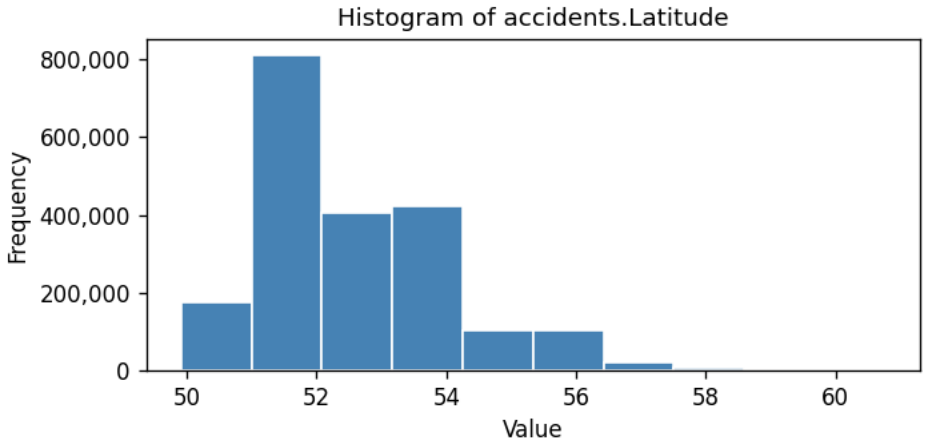
| Dimension | Issues |
|---|---|
| ACCURACY | 2 |
| COMPLETENESS | 3 |
| JOIN_INTEGRITY | 2 |
| WEATHER_DISTRIBUTION | 11 |

## Correlation Heatmap

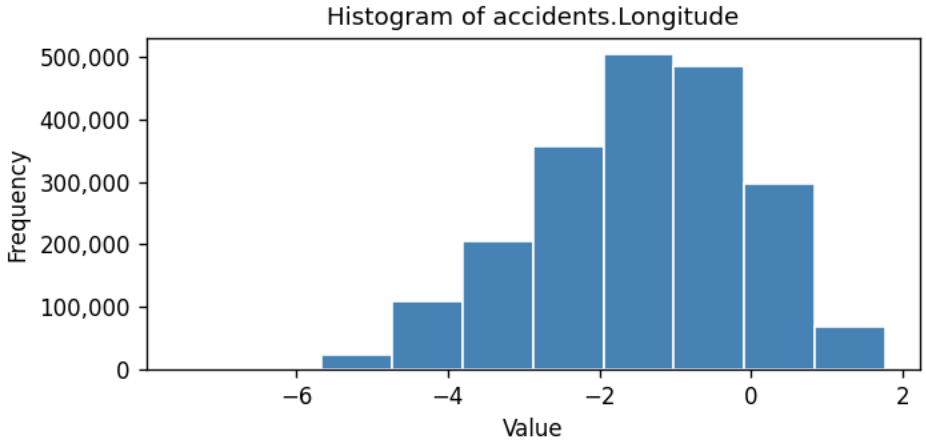


## Distribution Histograms

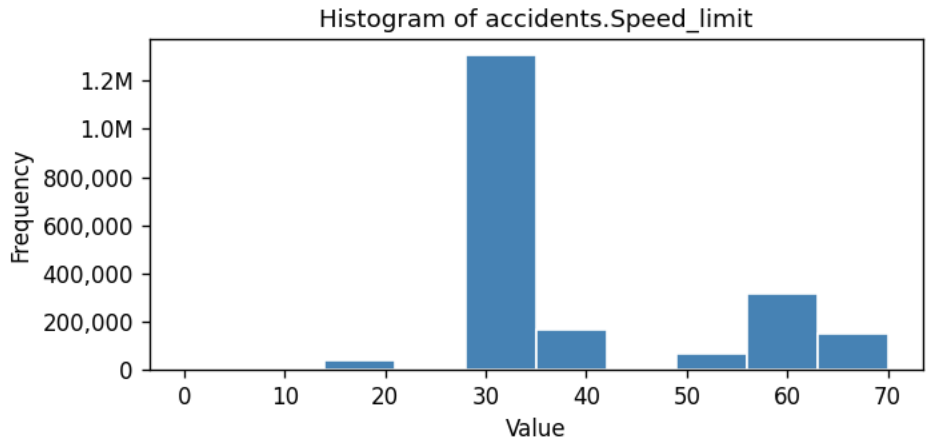
### *accidents.Latitude*



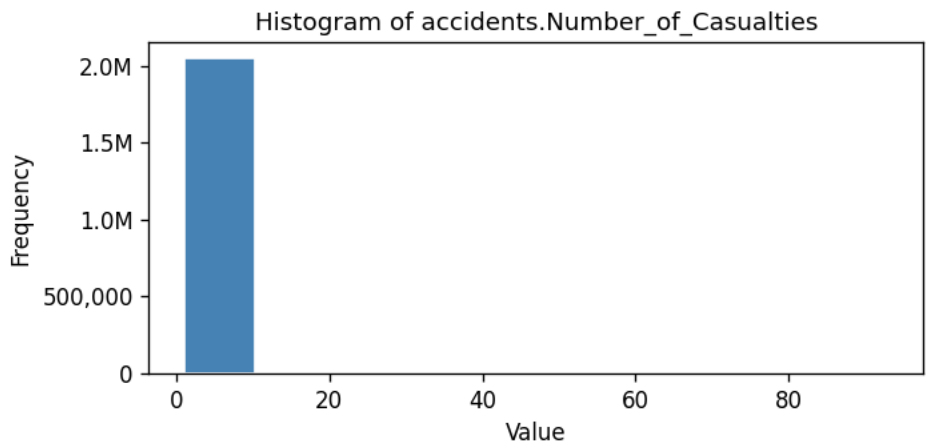

### *accidents.Longitude*



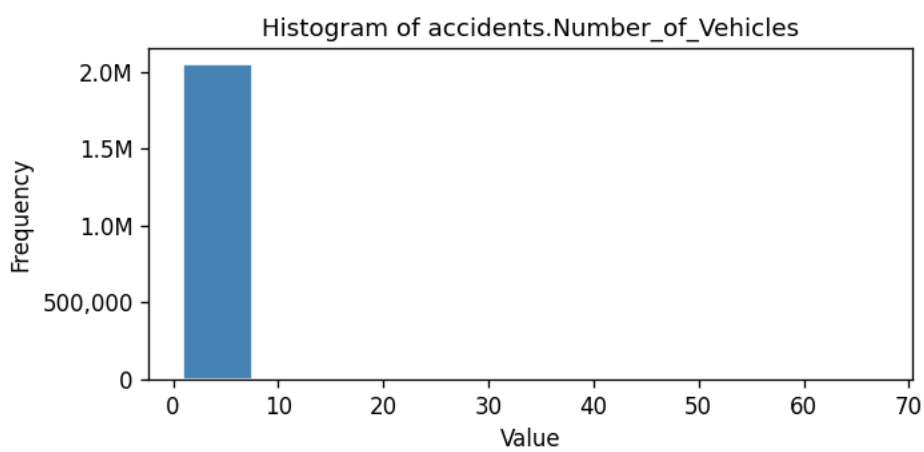

### *accidents.Speed_limit*



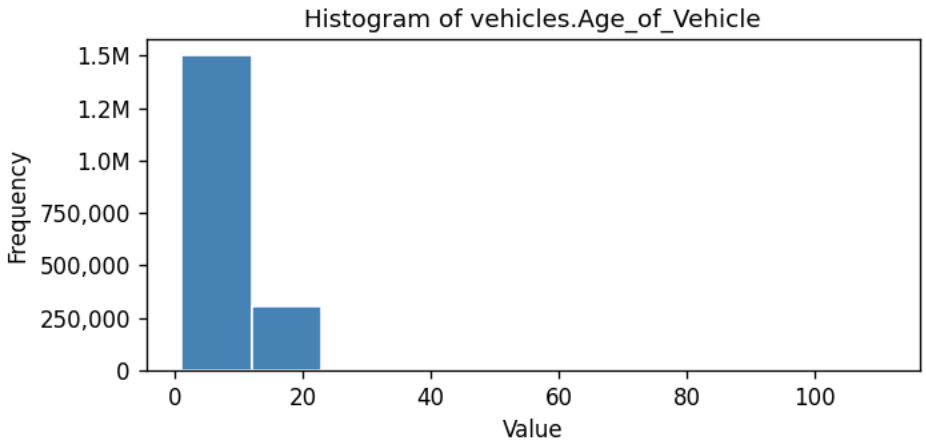

***accidents.Number_of_Casualties***



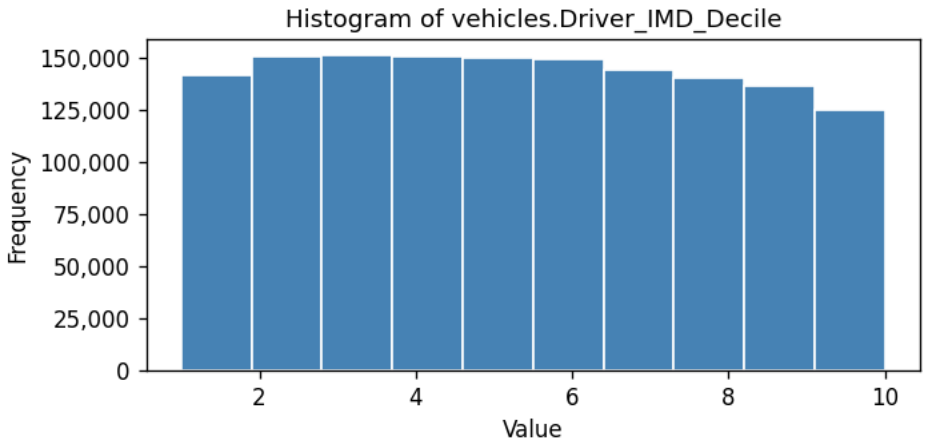

***accidents.Number_of_Vehicles***



***vehicles.Age_of_Vehicle***




***vehicles.Driver_IMD_Decile***



# Full Validation Log

```
================================================================
UK Road Accidents — Data Validation Report
2026-05-02 05:35:20
================================================================
Loaded accidents : 2,047,256 rows x 34 columns
Loaded vehicles : 2,177,205 rows x 24 columns
----------------------------------------------------------------
DIMENSION 1 — ACCURACY
----------------------------------------------------------------
[PASS] Accident_Severity — all values valid (Slight/Serious/Fatal)
[ISSUE] 36 rows have Speed_limit not in UK legal set {50, 20, 70, 40, 60, 30}
[PASS] Latitude — all within UK bounds (49–61)
[PASS] Longitude — all within UK bounds (-8 to 2)
[PASS] Number_of_Vehicles — all values >= 1
[PASS] Age_of_Vehicle — no negative values
[INFO] Age_Band_of_Driver has 12 unique categories
Sample values: ['26 - 35', '66 - 75', '36 - 45', '46 - 55', '21 - 25', 'Data missing or out of range', '16 - 20', '
56 - 65', 'Over 75', '11 - 15']
[PASS] Day_of_Week — all values valid
[ISSUE] 76,119 vehicle records have invalid Sex_of_Driver values
----------------------------------------------------------------
DIMENSION 2 — CONSISTENCY
----------------------------------------------------------------
[PASS] Date column — parses correctly as datetime
[PASS] Year column — matches year extracted from Date in all rows
[WARN] 2 Date/Day_of_Week mismatches found
[PASS] Time column — all values follow HH:MM format
[PASS] Year — consistent across both tables for all matched accidents
----------------------------------------------------------------
DIMENSION 3 — COMPLETENESS
----------------------------------------------------------------
[ACCIDENTS]
Column Missing %
------------------------------------------- --------- ------
Carriageway_Hazards 2,007,807 98.1%
Special_Conditions_at_Site 1,995,137 97.5%
2nd_Road_Class 844,272 41.2%
LSOA_of_Accident_Location 144,953 7.1%
2nd_Road_Number 17,593 0.9%
Pedestrian_Crossing-Physical_Facilities 3,560 0.2%
Pedestrian_Crossing-Human_Control 2,920 0.1%
Did_Police_Officer_Attend_Scene_of_Accident 278 0.0%
Latitude 174 0.0% [!] REQUIRED
Location_Easting_OSGR 164 0.0%
Location_Northing_OSGR 164 0.0%
Longitude 175 0.0% [!] REQUIRED
[ISSUE] [ACCIDENTS] CRITICAL: 'Latitude' has 174 nulls (0.01%). These rows must be dropped.
[ISSUE] [ACCIDENTS] CRITICAL: 'Longitude' has 175 nulls (0.01%). These rows must be dropped.
[ISSUE] [ACCIDENTS] CRITICAL: 'Time' has 156 nulls (0.01%). These rows must be dropped.
[ADVICE] 'Carriageway_Hazards' is 98.07% missing. Consider dropping this feature.
[ADVICE] 'Special_Conditions_at_Site' is 97.45% missing. Consider dropping this feature.
[VEHICLES]
Column Missing %
------------------------------------------- --------- ------
Hit_Object_in_Carriageway 2,087,824 95.9%
Hit_Object_off_Carriageway 1,989,741 91.4%
Skidding_and_Overturning 1,898,208 87.2%
Driver_IMD_Decile 734,812 33.8%
Age_of_Vehicle 358,149 16.4%
model 325,331 14.9%
Engine_Capacity_.CC. 265,861 12.2%
Propulsion_Code 245,843 11.3%
make 110,845 5.1%
Vehicle_Location.Restricted_Lane 1,317 0.1%
[PASS] All strictly required columns in VEHICLES are 100% complete.
[ADVICE] 'Hit_Object_in_Carriageway' is 95.89% missing. Consider dropping this feature.
```

```
[ADVICE] 'Hit_Object_off_Carriageway' is 91.39% missing. Consider dropping this feature.
[ADVICE] 'Skidding_and_Overturning' is 87.19% missing. Consider dropping this feature.
----------------------------------------------------------------
DIMENSION 4 — UNIQUENESS (POST-MERGE)
----------------------------------------------------------------
[PASS] Zero fully duplicate rows in merged data.
[PASS] Composite key (Accident_Index, Vehicle_Reference) is unique.
[INFO] Found 668,684 repeated Accident_Indices (Expected: multiple vehicles per accident).
----------------------------------------------------------------
DIMENSION 5 — OUTLIERS
----------------------------------------------------------------
IQR Method (bounds = Q1 - 1.5*IQR and Q3 + 1.5*IQR):
Column Outliers Lower fence Upper fence
---------------------------------------- ---------- ------------ ------------
Isolation Forest (contamination=0.05, multivariate):
accidents: 4,858 anomalies in 100,000 rows (4.9%)
vehicles: 4,999 anomalies in 100,000 rows (5.0%)
[NOTE] Outliers are FLAGGED only — removal decisions made in cleaning step.
----------------------------------------------------------------
DIMENSION 6 — TIMELINESS
----------------------------------------------------------------
Earliest record : 2005-01-01
Latest record : 2017-12-31
[PASS] No future dates
[PASS] No dates before 2005
----------------------------------------------------------------
DIMENSION 7 — DISTRIBUTION PROFILE
----------------------------------------------------------------
Column Min Max Mean Median Std Skew Kurt Card
---------------------------------------- ------- -------- -------- -------- -------- ------- ------- ------
accidents.Latitude 49.9 60.8 52.56 52.24 1.45 1.03 0.83 1,319,367
Histogram (accidents.Latitude):
[ 49.91 – 51.0] 175,171 ( 8.6%) ■■■■
[ 51.0 – 52.08] 809,444 ( 39.5%) ■■■■■■■■■■■■■■■■■■■■
[ 52.08 – 53.17] 405,212 ( 19.8%) ■■■■■■■■■
[ 53.17 – 54.25] 424,012 ( 20.7%) ■■■■■■■■■■
[ 54.25 – 55.34] 102,546 ( 5.0%) ■■
[ 55.34 – 56.42] 102,828 ( 5.0%) ■■
[ 56.42 – 57.5] 20,906 ( 1.0%) ■
[ 57.5 – 58.59] 6,073 ( 0.3%) ■
[ 58.59 – 59.67] 492 ( 0.0%) ■
[ 59.67 – 60.76] 398 ( 0.0%) ■
accidents.Longitude -7.5 1.8 -1.41 -1.36 1.40 -0.39 -0.35 1,417,728
Histogram (accidents.Longitude):
[ -7.52 – -6.59] 225 ( 0.0%) ■
[ -6.59 – -5.66] 1,071 ( 0.1%) ■
[ -5.66 – -4.73] 23,534 ( 1.1%) ■
[ -4.73 – -3.8] 108,905 ( 5.3%) ■■
[ -3.8 – -2.88] 204,273 ( 10.0%) ■■■■■
[ -2.88 – -1.95] 356,689 ( 17.4%) ■■■■■■■■
[ -1.95 – -1.02] 503,647 ( 24.6%) ■■■■■■■■■■■■
[ -1.02 – -0.09] 484,383 ( 23.7%) ■■■■■■■■■■■
[ -0.09 – 0.83] 296,607 ( 14.5%) ■■■■■■■
[ 0.83 – 1.76] 67,747 ( 3.3%) ■
accidents.Speed_limit 0.0 70.0 38.84 30.00 14.15 1.10 -0.36 9
Histogram (accidents.Speed_limit):
[ 0.0 – 7.0] 1 ( 0.0%) ■
[ 7.0 – 14.0] 19 ( 0.0%) ■
[ 14.0 – 21.0] 38,415 ( 1.9%) ■
[ 21.0 – 28.0] 0 ( 0.0%) ■
[ 28.0 – 35.0] 1,306,174 ( 63.8%) ■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■
[ 35.0 – 42.0] 168,357 ( 8.2%) ■■■■
[ 42.0 – 49.0] 0 ( 0.0%) ■
[ 49.0 – 56.0] 69,479 ( 3.4%) ■
[ 56.0 – 63.0] 317,469 ( 15.5%) ■■■■■■■
[ 63.0 – 70.0] 147,305 ( 7.2%) ■■■
accidents.Number_of_Casualties 1.0 93.0 1.35 1.00 0.82 8.57 412.75 52
Histogram (accidents.Number_of_Casualties):
[ 1.0 – 10.2] 2,046,795 (100.0%) ■■■■■■■■■■■■■■■■■■■
```

```
[ 19.4 – 28.6] 56 ( 0.0%) ■
[ 28.6 – 37.8] 17 ( 0.0%) ■
[ 37.8 – 47.0] 15 ( 0.0%) ■
[ 47.0 – 56.2] 5 ( 0.0%) ■
[ 56.2 – 65.4] 3 ( 0.0%) ■
[ 65.4 – 74.6] 2 ( 0.0%) ■
[ 74.6 – 83.8] 0 ( 0.0%) ■
[ 83.8 – 93.0] 3 ( 0.0%) ■
accidents.Number_of_Vehicles 1.0 67.0 1.83 2.00 0.71 2.16 53.08 29
Histogram (accidents.Number_of_Vehicles):
[ 1.0 – 7.6] 2,046,478 (100.0%) ■■■
```

```
[ 7.35 – 11.46] 351,669 ( 25.3%) ■■■■■■■■■■■■■
[ 11.46 – 15.57] 383,705 ( 27.6%) ■■■■■■■■■■■■■■
[ 15.57 – 19.68] 226,063 ( 16.3%) ■■■■■■■■
[ 19.68 – 23.79] 36,972 ( 2.7%) ■
[ 23.79 – 27.90] 3,274 ( 0.2%) ■
KS=0.0464 p=0.000000 → NOT normal (p<0.05) — use Spearman
[ISSUE] [WEATHER] 'tmin' is 56.8% missing (>20.0% threshold)
tmin 56.8% -24.20 26.10 7.12 7.40 5.20 -0.24 -0.51
[PASS] tmin — all values within physical bounds [-30, 45]
Histogram (tmin):
[ -24.20 – -19.17] 15 ( 0.0%) ■
[ -19.17 – -14.14] 79 ( 0.0%) ■
[ -14.14 – -9.11] 879 ( 0.1%) ■
[ -9.11 – -4.08] 14,275 ( 1.2%) ■
[ -4.08 – 0.95] 137,593 ( 11.7%) ■■■■■
[ 0.95 – 5.98] 307,622 ( 26.2%) ■■■■■■■■■■■■■■
[ 5.98 – 11.01] 422,566 ( 36.0%) ■■■■■■■■■■■■■■■■■■■
[ 11.01 – 16.04] 268,183 ( 22.9%) ■■■■■■■■■■■
[ 16.04 – 21.07] 22,117 ( 1.9%) ■
[ 21.07 – 26.10] 190 ( 0.0%) ■
KS=0.0585 p=0.000000 → NOT normal (p<0.05) — use Spearman
[ISSUE] [WEATHER] 'tmax' is 56.7% missing (>20.0% threshold)
tmax 56.7% -8.00 36.70 14.22 14.10 5.88 0.04 -0.40
[PASS] tmax — all values within physical bounds [-25, 50]
Histogram (tmax):
[ -8.00 – -3.53] 252 ( 0.0%) ■
[ -3.53 – 0.94] 5,819 ( 0.5%) ■
[ 0.94 – 5.41] 75,094 ( 6.4%) ■■■
[ 5.41 – 9.88] 201,948 ( 17.2%) ■■■■■■■■
[ 9.88 – 14.35] 314,318 ( 26.7%) ■■■■■■■■■■■■■■
[ 14.35 – 18.82] 299,756 ( 25.5%) ■■■■■■■■■■■■■
[ 18.82 – 23.29] 211,476 ( 18.0%) ■■■■■■■■■
[ 23.29 – 27.76] 54,974 ( 4.7%) ■■
[ 27.76 – 32.23] 11,470 ( 1.0%) ■
[ 32.23 – 36.70] 603 ( 0.1%) ■
KS=0.0460 p=0.000000 → NOT normal (p<0.05) — use Spearman
[ISSUE] [WEATHER] 'prcp' is 94.5% missing (>20.0% threshold)
prcp 94.5% 0.00 221.50 2.01 0.30 4.44 6.48 147.83
[PASS] prcp — all values within physical bounds [0, 10000]
Histogram (prcp):
[ 0.00 – 22.15] 148,338 ( 99.2%) ■■■■■■■■■■■■■■■■■■■
```

```
[ 0.00 – 9.09] 222,243 ( 16.9%) ■■■■■■■■
[ 9.09 – 18.18] 696,070 ( 53.1%) ■■■■■■■■■■■■■■■■■■■■■■■■■■■■■
[ 18.18 – 27.27] 307,303 ( 23.4%) ■■■■■■■■■■■■
[ 27.27 – 36.36] 70,111 ( 5.3%) ■■
[ 36.36 – 45.45] 12,311 ( 0.9%) ■
[ 45.45 – 54.54] 2,685 ( 0.2%) ■
[ 54.54 – 63.63] 676 ( 0.1%) ■
[ 63.63 – 72.72] 341 ( 0.0%) ■
[ 72.72 – 81.81] 198 ( 0.0%) ■
[ 81.81 – 90.90] 88 ( 0.0%) ■
KS=0.0735 p=0.000000 → NOT normal (p<0.05) — use Spearman
[ISSUE] [WEATHER] 'pres' is 73.9% missing (>20.0% threshold)
pres 73.9% 955.20 1044.20 1014.84 1015.70 10.78 -0.46 0.43
[PASS] pres — all values within physical bounds [870, 1085]
Histogram (pres):
[ 955.20 – 964.10] 13 ( 0.0%) ■
[ 964.10 – 973.00] 439 ( 0.1%) ■
[ 973.00 – 981.90] 2,643 ( 0.4%) ■
[ 981.90 – 990.80] 14,660 ( 2.1%) ■
[ 990.80 – 999.70] 45,697 ( 6.4%) ■■■
[ 999.70 – 1008.60] 113,763 ( 16.0%) ■■■■■■■■
[ 1008.60 – 1017.50] 227,890 ( 32.1%) ■■■■■■■■■■■■■■■■
[ 1017.50 – 1026.40] 213,322 ( 30.1%) ■■■■■■■■■■■■■■■
[ 1026.40 – 1035.30] 77,405 ( 10.9%) ■■■■■
[ 1035.30 – 1044.20] 13,755 ( 1.9%) ■
KS=0.0416 p=0.000000 → NOT normal (p<0.05) — use Spearman
[ISSUE] [WEATHER] 'rhum' is 49.4% missing (>20.0% threshold)
rhum 49.4% 4.00 100.00 79.99 81.00 9.81 -0.44 -0.10
[PASS] rhum — all values within physical bounds [0, 100]
Histogram (rhum):
[ 4.00 – 13.60] 4 ( 0.0%) ■
[ 13.60 – 23.20] 0 ( 0.0%) ■
[ 23.20 – 32.80] 0 ( 0.0%) ■
[ 32.80 – 42.40] 200 ( 0.0%) ■
[ 42.40 – 52.00] 5,274 ( 0.4%) ■
[ 52.00 – 61.60] 50,840 ( 3.7%) ■
[ 61.60 – 71.20] 212,181 ( 15.4%) ■■■■■■■
[ 71.20 – 80.80] 402,660 ( 29.3%) ■■■■■■■■■■■■■■■
[ 80.80 – 90.40] 510,567 ( 37.2%) ■■■■■■■■■■■■■■■■■■■
[ 90.40 – 100.00] 192,139 ( 14.0%) ■■■■■■■
KS=0.0652 p=0.000000 → NOT normal (p<0.05) — use Spearman
[ISSUE] [WEATHER] 'wpgt' is 100.0% missing (>20.0% threshold)
wpgt 100.0% [SKIP — too few valid values]
[ISSUE] [WEATHER] 'tsun' is 100.0% missing (>20.0% threshold)
tsun 100.0% [SKIP — too few valid values]
[ISSUE] [WEATHER] 'cldc' is 85.6% missing (>20.0% threshold)
cldc 85.6% 0.00 8.00 5.09 5.00 1.80 -0.71 0.23
[PASS] cldc — all values within physical bounds [0, 9]
Histogram (cldc):
[ 0.00 – 0.80] 9,133 ( 2.3%) ■
[ 0.80 – 1.60] 8,549 ( 2.2%) ■
[ 1.60 – 2.40] 18,176 ( 4.6%) ■■
[ 2.40 – 3.20] 29,460 ( 7.5%) ■■■
[ 3.20 – 4.00] 0 ( 0.0%) ■
[ 4.00 – 4.80] 68,104 ( 17.4%) ■■■■■■■■
[ 4.80 – 5.60] 71,745 ( 18.3%) ■■■■■■■■■
[ 5.60 – 6.40] 96,445 ( 24.6%) ■■■■■■■■■■■■
[ 6.40 – 7.20] 69,912 ( 17.9%) ■■■■■■■■
[ 7.20 – 8.00] 19,925 ( 5.1%) ■■
KS=0.1647 p=0.000000 → NOT normal (p<0.05) — use Spearman
----------------------------------------------------------------
DIMENSION 8 — RELATIONSHIPS
----------------------------------------------------------------
Pearson Correlation Matrix:
Speed_limit Number_of_C Number_of_V Latitude Longitude
Speed_limit 1.000 0.143 0.088 0.052 -0.066
Number_of_Casualties 0.143 1.000 0.237 0.030 -0.038
Number_of_Vehicles 0.088 0.237 1.000 -0.039 0.019
Latitude 0.052 0.030 -0.039 1.000 -0.434
Longitude -0.066 -0.038 0.019 -0.434 1.000
```

```
Pearson vs Spearman (|diff| > 0.10 = non-linear relationship detected):
Pair Pearson Spearman |Diff| Note
------------------------------------------ --------- --------- -------- -------------------------
Speed_limit x Number_of_Casualties 0.143 0.162 0.019
Speed_limit x Number_of_Vehicles 0.088 0.069 0.019
Speed_limit x Latitude 0.052 0.034 0.018
Speed_limit x Longitude -0.066 -0.062 0.004
Number_of_Casualties x Number_of_Vehicles 0.237 0.230 0.007
Number_of_Casualties x Latitude 0.030 0.044 0.014
Number_of_Casualties x Longitude -0.038 -0.046 0.008
Number_of_Vehicles x Latitude -0.039 -0.029 0.010
Number_of_Vehicles x Longitude 0.019 0.018 0.001
Latitude x Longitude -0.434 -0.411 0.023
[PASS] No high-correlation pairs found (threshold |r| > 0.85)
[SAVED] reports\correlation_heatmap.png
----------------------------------------------------------------
JOIN INTEGRITY (Accident_Index key across both tables)
----------------------------------------------------------------
Unique Accident_Index in accidents table : 2,047,256
Unique Accident_Index in vehicles table : 1,488,981
Keys matched (will survive inner join) : 1,389,724
[ISSUE] 657,532 accidents (32.1%) have no vehicle record — dropped on inner join, document in cleaning step
[ISSUE] 99,257 vehicle records (6.7%) have no matching accident — orphaned records, dropped on inner join
================================================================
VALIDATION COMPLETE
Total issues found across all 8 dimensions: 18
================================================================
Issues by dimension:
[ACCURACY] 36 rows have Speed_limit not in UK legal set {50, 20, 70, 40, 60, 30}
[ACCURACY] 76,119 vehicle records have invalid Sex_of_Driver values
[COMPLETENESS] [ACCIDENTS] CRITICAL: 'Latitude' has 174 nulls (0.01%). These rows must be dropped.
[COMPLETENESS] [ACCIDENTS] CRITICAL: 'Longitude' has 175 nulls (0.01%). These rows must be dropped.
[COMPLETENESS] [ACCIDENTS] CRITICAL: 'Time' has 156 nulls (0.01%). These rows must be dropped.
[WEATHER_DISTRIBUTION] [WEATHER] 'temp' is 48.8% missing (>20.0% threshold)
[WEATHER_DISTRIBUTION] [WEATHER] 'tmin' is 56.8% missing (>20.0% threshold)
[WEATHER_DISTRIBUTION] [WEATHER] 'tmax' is 56.7% missing (>20.0% threshold)
[WEATHER_DISTRIBUTION] [WEATHER] 'prcp' is 94.5% missing (>20.0% threshold)
[WEATHER_DISTRIBUTION] [WEATHER] 'snwd' is 98.7% missing (>20.0% threshold)
[WEATHER_DISTRIBUTION] [WEATHER] 'wspd' is 51.7% missing (>20.0% threshold)
[WEATHER_DISTRIBUTION] [WEATHER] 'pres' is 73.9% missing (>20.0% threshold)
[WEATHER_DISTRIBUTION] [WEATHER] 'rhum' is 49.4% missing (>20.0% threshold)
[WEATHER_DISTRIBUTION] [WEATHER] 'wpgt' is 100.0% missing (>20.0% threshold)
[WEATHER_DISTRIBUTION] [WEATHER] 'tsun' is 100.0% missing (>20.0% threshold)
[WEATHER_DISTRIBUTION] [WEATHER] 'cldc' is 85.6% missing (>20.0% threshold)
[JOIN_INTEGRITY] 657,532 accidents (32.1%) have no vehicle record — dropped on inner join, document in cleaning step
[JOIN_INTEGRITY] 99,257 vehicle records (6.7%) have no matching accident — orphaned records, dropped on inner join
[SAVED] reports\validation_report.json
[SAVED] reports\validation_summary.txt
```