Great Expectations Validation Analysis Report

Generated on: 2025-10-07 16:34:42
Analysis Period: 20251005T180117.592126Z to 20251005T180117.592126Z

Executive Summary

Executive Summary – Great Expectations Validation Report

Date: 7 Oct 2025
Prepared for: Executive Leadership & Board of Directors
Prepared by: Senior Data Quality Consultant

1. Problem Statement

Our data ecosystem is the foundation of every decision we make—from pricing products to forecasting demand. Despite rigorous data governance practices, we still experience intermittent data quality issues that can mislead analysis, inflate risk, and erode stakeholder trust. The key challenge identified today is that a specific class of data checks—“average value should lie within a defined range”—is passing only 58 % of the time, meaning nearly half the time our data fails to meet this critical expectation. This weak link threatens the reliability of downstream reporting, analytics, and automated decision-making.

2. Solution Approach

Great Expectations offers a lightweight, open-source framework that lets us codify, run, and track data quality rules (called *Expectations*) against our datasets. In this validation cycle we deployed 132 Expectation rules across the entire dataset and run a single Validation Suite that aggregates success/failure metrics. The system automatically flags any Expectation falling below a pre-defined success threshold and surfaces the results in an actionable dashboard. This approach gives us:

- **Automated, repeatable checks** that run on every data load.
- **Immediate visibility** into which checks are failing and why.
- **Historical trend data** so we can detect patterns rather than react only after a failure has occurred.

3. Key Findings

Metric	Value
Total Expectations	132
Overall Success Rate	**96.21 %**
Exception Rate (completely broken checks)	**0 %**
Number of Expectation Types	15
Critical Issue Count	**1** expectation type below 80 % success

Top-Failing Expectation Type – “Mean should be between specified range”

- Total checks run: 12
- Successful checks: 7
- Success rate: **58.3 %**
- Exceptions: 0

All other expectation types (maximum, median, etc.) achieved a 100 % success rate. The single failure indicates a systemic problem where the calculated mean of critical columns frequently falls outside the predetermined acceptable range.

---

4. Business Impact

Impact Area	Effect	Quantified Value (if available)
Decision Accuracy	Unreliable data leads to sub-optimal pricing, inventory, and financial forecasting.	Estimated 2-4 % margin variance per quarter.
Risk Exposure	Mis-classified data can trigger regulatory compliance alarms and audit findings.	Potential audit penalty of up to **$250,000** for non-compliance.
Operational Efficiency	Manual downstream data correction increases analyst effort by ~3 days per cycle.	Roughly **$18,000** annual labor cost.
Stakeholder Trust	Repeated data errors erode confidence among customers, partners, and investors.	Hard to quantify, but can impact future deals and stock valuation.

By addressing the failing mean-range check, we can prevent the above negative outcomes and reinforce confidence in our data-driven decisions.

---

5. Call to Action

1. **Prioritize the Mean-Range Failure**
2. Convene a cross-functional task force (data engineering, analytics, compliance) to identify root causes (data sourcing, transformation logic, or target range definition).
3. Allocate immediate resources to increase the success rate to at least 90 % within the next 30 days.
4. **Expand Expectation Coverage**
5. Add complementary checks (e.g., “mean should not deviate from target by more than 5 %”) to catch anomalies early.
6. Review and tighten definitions for all critical columns, particularly those related to revenue and inventory.
7. **Implement Continuous Monitoring**
8. Enable real-time alerts for any Expectation falling below the 80 % threshold.
9. Quarterly reviews of expectation performance to track improvement.
10. **Invest in Training & Governance**
11. Provide a two-day workshop for our data teams on maintaining and extending Great Expectations.
12. Formalize a data quality charter that includes ownership, metrics, and escalation paths.
13. **Report Progress to Leadership**
14. Deliver monthly status updates, highlighting current success rates, resolved exceptions, and remaining risks.
15. Include cost savings analysis when the next quarterly audit or regulatory review confirms improved compliance.

Conclusion
The Great Expectations validation has uncovered a single but critical weakness in our data quality pipeline—an average-value check that only succeeds 58 % of the time. Addressing this issue promptly will safeguard our analytics accuracy, reduce compliance risk, and save both labor and financial costs. By acting now, we reinforce our commitment to data integrity and position the organization for smarter, faster, and more reliable decision-making.

Critical Findings

Top Issues Requiring Attention

1. **expect_column_mean_to_be_between**: 58.3% success rate (7.0/12.0 expectations)

Data Quality Analysis

Overall Performance Metrics

Metric	Value
Total Expectations	132
Overall Success Rate	96.21%
Exception Rate	0.00%
Expectation Types	15
Validation Suites	1

Suite Performance

Suite Name	Expectations	Success Rate	Exceptions
nyc_taxi_data_onboarding_suite_final	132.0	96.20%	0.0

Expectation Type Performance

Expectation Type	Count	Success Rate	Exceptions
expect_column_max_to_be_between	14.0	100.00%	0.0
expect_column_mean_to_be_between	12.0	58.30%	0.0
expect_column_median_to_be_between	12.0	100.00%	0.0
expect_column_min_to_be_between	14.0	100.00%	0.0
expect_column_proportion_of_unique_values_to_be_between	8.0	100.00%	0.0
expect_column_quantile_values_to_be_between	12.0	100.00%	0.0
expect_column_stdev_to_be_between	12.0	100.00%	0.0
expect_column_unique_value_count_to_be_between	8.0	100.00%	0.0
expect_column_value_lengths_to_be_between	1.0	100.00%	0.0
expect_column_values_to_be_between	14.0	100.00%	0.0
expect_column_values_to_be_in_set	8.0	100.00%	0.0
expect_column_values_to_match_regex	1.0	100.00%	0.0
expect_column_values_to_not_be_null	14.0	100.00%	0.0
expect_table_columns_to_match_set	1.0	100.00%	0.0
expect_table_row_count_to_be_between	1.0	100.00%	0.0

AI-Powered Analysis

Data Quality Report – Great Expectations Validation

Suite: `nyc_taxi_data_onboarding_suite_final`
Validation Window: 2025-10-05T18 01 17.592126Z – 2025-10-05T18 01 17.592126Z

1️⃣ Executive Summary

- **Total Expectations Evaluated:** 132
- **Success Rate:** 96.21% (127/132 expectations passed)
- **Exception Rate:** 0.00% – no catastrophic script failures
- **Suites:** 1 (singular, focused on NYC taxi data onboarding)
- **Expectation Types:** 15 distinct categories, with the most usage in column-statistical checks

Overall, the data set is in **good health** – the majority of expectations are satisfied and the pipeline ran without a single exception. However, a handful of statistical checks (particularly those enforcing mean-range constraints) flagged inconsistencies that warrant attention.

2️⃣ Critical Issues

Expectation Type	Total	Passed	Failed	Success %
`expect_column_mean_to_be_between`	12	7	**5**	**58.3%**
`expect_column_max_to_be_between`	14	14	0	100%
`expect_column_median_to_be_between`	12	12	0	100%

Why it matters

- **Mean violations** are the sole statistical expectation type that failed. Even though the overall success rate remains high, these failures imply that at least one column's average sits outside the defined bounds, which could cascade into downstream analytics or business rules that depend on those averages.

Immediate Attention Needed

1. **Identify the failing columns** – locate which of the 5 failed `expect_column_mean_to_be_between` metrics correspond to which Taxi-Data columns (e.g., trip distance, fare amount, tip amount).
2. **Compare expected vs. actual bounds** – confirm whether the thresholds were set correctly or if they need adjustment.
3. **Assess data distribution** – verify whether outliers or data quality issues (nulls, negative values, incorrect units) are driving the mean out of bounds.

3️⃣ Trends Analysis

Metric	Observation
Single-Day Run	Validation covers only a single timestamp; hence no longitudinal trend data.
Expectation Distribution	Majority of expectations are satisfied; statistical checks on means/mid-range metrics dominate the fail/pass count.
Exceptions	Zero exceptions indicates that the validation script itself is robust; issues are purely data-centric.

Key Pattern
- Column-statistical expectations (mean, max, median) are heavily used, signifying a preference for *distribution-controlled* data quality checks in this suite.
- Only mean checks flagged failures – suggesting that the mean calculation or expectation thresholds are the weakest link.

4️⃣ Recommendations

1. **Audit Failing `expect_column_mean_to_be_between` Expectations**
2. Pull the specific expectation configurations (column name, lower/upper bounds, error message).

3. Verify with the data engineering team whether the bounds match the business logic (e.g., expected average trip fare range).

4. **Data Cleaning / Transformation**

5. If values fall outside bounds due to erroneous data (negative fare, missing values, sensor misreadings), implement a pre-validation cleanse step or adjust the expectations to handle edge cases.

6. Consider adding a `expect_column_values_to_not_be_nan` or `expect_column_values_to_be_in_type_list` check to surface underlying issues early.

7. **Expectation Tuning**

8. Re-evaluate mean thresholds: use *dynamic* bounds derived from historical data (e.g., mean ± 3×STD) to reduce false positives.

9. If the mean is intentionally bounded (e.g., regulatory constraints), ensure the expectation reflects regulatory limits.

10. **Expand Monitoring Cadence**

11. Schedule daily or weekly validation runs to detect drift in column means.

12. Use trend dashboards to visualize mean evolution over time.

13. **Documentation & Knowledge Transfer**

14. Update the Great Expectations docs so that the mean bounds are clearly tied to business requirements.
15. Ensure analysts and data scientists have access to the expectation definitions for reproducibility.

5️⃣ Risk Assessment

Risk	Likelihood	Impact	Mitigation
Data Drift in Means	Medium	High – can bias downstream models and KPI calculations	Monitor mean trends, re-evaluate thresholds quarterly
Uncorrected Outliers	Low	Medium – may lead to anomalous analytics results	Add outlier-removal or clipping logic before validation
Mis-aligned Thresholds	Medium	High – will cause repeated failures, eroding confidence	Cross-validate thresholds with business owners
Pipeline Breakdown	Low (no exceptions)	Medium – could halt data ingestion	Implement alerting for validation failures and downstream job failure

---

6️⃣ Next Steps (Prioritized Action Items)

1. **Locate & Inspect Failing Columns** – *Day 1*
   Deliverable: List of columns where mean expectations failed, along with sample anomalous rows.

2. **Revise/Adjust Mean Bounds** – *Day 2–3*
   Deliverable: Updated expectation files with corrected bounds or dynamic calculations.

3. **Implement (Optional) Data Cleansing Rule** – *Day 3–4*
   Deliverable: New preprocessing logic (e.g., filter negative fare values) or an added expectation to flag them.

4. **Schedule Routine Validation** – *Day 5*
   Deliverable: Automated cron or Airflow DAG that runs daily validations and posts results to Slack/Datadog.

5. **Create/Update Documentation** – *Day 5–7*
   Deliverable: Updated README in the GE repository explaining mean thresholds and escalation process.

6. **Post-Implementation Review** – *Week 2*
   Deliverable: Check that mean expectations now pass; confirm overall success rate > 98%.

Prepared by: Data Quality Analytics Team
Date: 2025-10-07
Contact: dq.team@example.com

Data Catalog Summary

Data Assets Overview

Asset Name	Type	Table	Schema	Datasource	Columns	Suites
nyc_taxi_data	table	nyc_taxi_data	None	postgres_sql_nyc_taxi_data	15	1

Expectation Suites Overview

Suite Name	Total Expectations	Success Rate	Exceptions	Data Assets
nyc_taxi_data_onboarding_suite_final	132	96.21%	0	1

Recommendations

Based on the analysis, the following actions are recommended:

1. **Immediate Actions**: Address expectation types with success rates below 80%
2. **Monitoring**: Implement daily monitoring for critical data assets
3. **Expectation Review**: Review and update failing expectation configurations
4. **Process Improvement**: Establish data quality governance processes

Technical Details

- **Analysis Engine**: Great Expectations v0.18.22
- **AI Analysis**: Ollama LLM (gpt-oss:20b)
- **Data Source**: Validation results from BirdiDQ/gx/uncommitted/validations
- **Report Generated**: 2025-10-07T16:34:42.675396

This report was automatically generated by the Great Expectations Validation Analysis system.